Modeling Transitivity in Complex Networks

Morteza Haghir Chehreghani*
Xerox Research Centre Europe - XRCE
6-8 Chemin de Maupertuis
38240 Meylan, France
morteza.chehreghani@xrce.xerox.com

Mostafa Haghir Chehreghani*
Department of Computer Science
KU Leuven
Leuven, 3001, Belgium
mostafa.chehreghani@gmail.com

Abstract

An important source of high clustering coefficient in real-world networks is *transitivity*. However, existing algorithms which model transitivity suffer from at least one of the following problems: *i*) they produce graphs of a specific class like bipartite graphs, *ii*) they do not give an analytical argument for the high clustering coefficient of the model, and *iii*) their clustering coefficient is still significantly lower than real-world networks. In this paper, we propose a new model for complex networks which is based on adding transitivity to scale-free models. We theoretically analyze the model and provide analytical arguments for its different properties. In particular, we calculate a lower bound on the clustering coefficient of the model which is independent of the network size, as seen in real-world networks. More than theoretical analysis, the main properties of the model are evaluated empirically and it is shown that the model can precisely simulate real-world networks from different domains with and different specifications.

1 Introduction

Most of real-world networks such as World Wide Web, social networks, Internet and biological networks exhibit structural properties which are not in either entirely regular or purely random graphs. For example, graphs produced by the model of Paul Erdős and Alfréd Rényi (the ER model) [10], do not have the two important properties observed in many real-world networks. The first property is related to the *degree distribution*. In a network, the *degree distribution* is defined as the probability distribution of the degrees of vertices over the whole network. In many real-world networks a *power-law distribution* is observed. More formally, the probability that the degree of a vertex is k is proportional to $k^{-\gamma}$. Networks with this property are called *scale-free* networks. However, the degree distribution of the graphs produced by the ER model converges to a *Poisson distribution*.

The second property is related to the *clustering coefficient*. Clustering coefficient is used to measure how well vertices in a network tend to be clustered together. In most of real-world networks, vertices tend to create tight groups characterized by dense ties [28]. However, in the ER model, every two vertices are connected with a constant and independent probability and therefore, the model generates graphs with a low clustering coefficient.

The β model (the Watts-Strogatz model), proposed by Watts and Strogatz [28], produces graphs with the *small-world* property and high *clustering coefficient*. In small-world networks, the distance between each pair of vertices is proportional to the logarithm of the number of vertices in the network. However, the β model produces an unrealistic degree distribution. The Barabási-Albert (BA) model, proposed by Albert-László Barabási and Réka Albert produces *scale-free* graphs [3]. The model is based on two important concepts: *growth* and *preferential attachment*. *Growth* means that the number of vertices in the network increases over time. *Preferential attachment* means that vertices with higher degree are more likely to receive new edges. The degree distribution of a graph resulting from the BA model is a power-law in the form of $\Pr[k] \sim k^{-3}$. However, the clustering coefficient of the graphs produced by the BA model is significantly lower than the clustering coefficient of real-world networks. Takemoto and Oosawa [25] propose a model for evolving networks by merging complete graphs (cliques) as building blocks. The model shows power-law degree distribution, power-law clustering spectra and high average clustering coefficients independent of the size of network. However, in most cases, real-world networks are evolved in a different way: they usually *grow* during the time by obtaining new vertices, rather than by merging complete graphs.

An important source of high clustering coefficient in net-

* The authors contributed equally.

works is *transitivity*. Transitivity means if u is connected to v and v is connected to w, the probability of having a connection between u and w is higher than any other pair of vertices in the network. Most of edges in real-world networks are *local* and they are drawn between vertices which have a common neighbor [18]. The model of [22] incorporates transitivity and generates graphs with high clustering coefficient. However, it produces bipartite networks which are limited to situations like company directors and movie actors. Clustering coefficient in the graphs produced by the model of [19] is still significantly lower than clustering coefficient of real-world networks. Leskovec et.al. [18] propose several mechanisms for modeling transitivity in complex networks. However, they do not provide any theoretical argument for the clustering coefficient of the mechanisms. The importance of such a theoretical analysis is that it guarantees that the model will reflect important properties of real-world networks, since a high clustering coefficient, independent of the network size, is seen in many real-world networks. On the other hand, for most of network models, it is not easy to theoretically analyze the clustering coefficient. For example, up to now, clustering coefficient of BA networks has only been determined by numerical simulations[1], and it is known to be very difficult to theoretically analyze it. Therefore, it is interesting to develop a model for transitivity in complex networks such that its clustering coefficient can be verified by theoretical arguments.

In this paper, we present the η model for modeling transitivity in complex networks. At every time interval t, the network obtains a new vertex and the new vertex is connected to some existing vertices. This step is similar to the BA model. Then, each vertex is selected with a probability proportional to its *degree* . If it is selected, then a pair of its neighbors are chosen randomly and an edge is drawn between them. The model has two adjustment parameter η and m. We theoretically analyze the model and prove that it produces networks with power-law degree distribution, high clustering coefficient and the small-world property. Compared to the clustering coefficient of random graphs or graphs produced by existing scale-free models, the clustering coefficient of the η model is significantly higher. In particular, by theoretical arguments, we prove that it is independent of the network size and depends solely on parameters like η and m. We also empirically evaluate the model and show that it can precisely simulate networks from different domains (biology, technology, social and information networks) with different characteristics.

The rest of this paper is organized as follows. In Second 2 we present the model and theoretically analyze its important properties. In Section 3 we empirically evaluate the model and show that it produces graphs very close to real-world networks. An overview of related work is given in Section 4, and finally the paper is concluded in Section 5.

[1]Numerical simulations show that clustering coefficient of a BA network with n vertices is $n^{-0.75}$.

Table 1: Symbols and their definitions.

Symbol	Definition
γ	The power-law exponent of the degree distribution in a scale-free network
η	A parameter of the proposed model ($\eta > 0$)
$\mathbf{G}$	A network produced by the η model
$V_{\mathbf{G}}$	The set of vertices of $\mathbf{G}$
$V_{\mathbf{G}}(t)$	The set of vertices of $\mathbf{G}$ at time t
$\mathbf{G}_0$	The initial graph
d_v	The degree of a vertex v
$d_v(t)$	The degree of a vertex v at time t
t_v	The time of adding v to the network
N_v	The set of neighbors of v
$N_v(t)$	The set of neighbors of v at time t
n	The number of vertices of the network
e	The number of edges of the network
$e(t)$	The number of edges of the network at time t
m	The number of edges drawn between a new vertex and the existing vertices of the network
$\langle CC \rangle$	The clustering coefficient
α	$\frac{2\eta+m}{2(\eta+m)}$
K	$\frac{2\eta}{2\eta+m}$

2 The η model

In this section, we first present the η model and then we theoretically analyze its important properties like power-law degree distribution, high clustering coefficient and the small-world property. Before that, in Table 1 we summarize symbols and notations that we will use in the paper.

Algorithm 1 describes the high level pseudo code of the η model proposed for modeling transitivity in complex networks. First a small graph $\mathbf{G}_0$ is produced. We refer to it as the *initial graph*. Then, at every time interval $t \in \{1, \ldots, \mathbf{T}\}$, the following steps are performed:
I. *growth.* A new vertex v is added to the network $\mathbf{G}$. We denote by t_v the time of adding v to $\mathbf{G}$.
II. *preferential attachment.* The vertex v is connected to m existing vertices. Existing vertices are chosen based on their degree. While every model which produces scale-free networks can be used, for the sake of simplicity, we here use the basic BA model. Therefore, for m times, a vertex w with probability

$$\frac{d_w(t)}{2e(t)} \tag{1}$$

is chosen and connected to v. We denote by $d_w(t)$ the degree of w at time interval t and by $e(t)$ the number of edges of the graph at time interval t.
III. *transitivity.* At this step each vertex w of the graph is

selected with probability

$$\frac{\eta d_w(t)}{2e(t)} \tag{2}$$

where η is a non-negative real number. Then, if w is selected, among the neighbors of w, two vertices are chosen *uniformly at random* and are connected to each other.

Algorithm 1 High level pseudo code of the η model.

GRAPHGENERATOR

Require: A non-negative real number η, a non-negative integer $\mathbf{T}$, a non-negative integer m.
Ensure: A graph $\mathbf{G}$ generated by the η model.

```
initialize G by a small graph
for t = 1 to T do
    {growth:}
    add a new vertex v to G
    {preferential attachment:}
    connect v to m existing vertices {every existing vertex is selected proportional to its degree}
    {transitivity:}
    for every vertex w ∈ V_G do
        select w with probability η d_w(t) / 2e(t)
        if w is selected then
            select two neighbors x and y of w uniformly at random
            draw an edge between x and y
        end if
    end for
end for
return G
```

The authors of [18] investigated different cases of producing triangles in complex networks. In their scenario, a source vertex u decides to connect to some vertex w whose distance with u is two. u first selects a neighbor v and then v selects a neighbor $w \neq u$. u and v might use different policies to select v and w, e.g. uniform selection or selecting based on degree. Here we first select v proportional to its degree and then, u and w are selected uniformly at random. The main contribution of this work compared to [18] is that we precisely formulate the procedure, which gives us a possibility to analytically study the model. Particularly, we provide a lower bound on the clustering coefficient independent of the network size.

2.1 Expected number of edges

In this section, we calculate the expected number of edges of the network at every time interval t.

The number of edges at time interval t, i.e. $e(t)$, satisfies the following dynamical equation:

$$\frac{\partial e(t)}{\partial t} = \underbrace{m}_{\text{preferential attachment}} + \underbrace{\sum_{w \in V_{\mathbf{G}}(t)} \frac{\eta d_w(t)}{2e(t)}}_{\text{transitivity}} = m + \eta$$

where $V_{\mathbf{G}}(t)$ denotes vertices of $\mathbf{G}$ at time interval t. After solving this equation, we obtain

$$e(t) = (m + \eta)t + e(\mathbf{G}_0) \tag{3}$$

where $e(\mathbf{G}_0)$ denotes the number of edges in the initial graph. For large enough t, we sometimes discard $e(\mathbf{G}_0)$ and consider $e(t)$ as $(m + \eta)t$.

2.2 Power-law degree distribution

In this section, we show that in a graph produced by the η model, vertices (except those added at the very early time intervals) have a power-law degree distribution.

At every time interval $t \in \{1, \dots, \mathbf{T}\}$, every vertex v in the network satisfies the following dynamical equation:

$$\begin{aligned}
\frac{\partial d_v(t)}{\partial t} &\overset{(a)}{\approx} \underbrace{\sum_{u \in N_v(t)} \left(\frac{\eta d_u(t)}{2e(t)} \times \frac{2}{d_u(t)} \right)}_{\text{transitivity}} + \underbrace{\frac{m d_v(t)}{2e(t)}}_{\text{preferential attachment}} \\
&= \sum_{u \in N_v(t)} \left(\frac{\eta}{e(t)} \right) + \frac{m d_v(t)}{2e(t)} \\
&= \frac{\eta d_v(t)}{e(t)} + \frac{m d_v(t)}{2e(t)}
\end{aligned} \tag{4}$$

where $N_v(t)$ refers to neighbors of vertex v at time interval t.

The approximation (a) is employed to make the computation of the dynamical equation $\frac{\partial d_v(t)}{\partial t}$ feasible, since, otherwise it would require taking the expectation of a function with a random variable at the denominator (i.e. the number of edges), which is computationally intractable. In principle, one could use the polynomial normal forms of such functions to eliminate the denominator. However, this transformation yields an exponential order in the number of conjunctions. Therefore, in mean-field theory, it is proposed to approximate the expectation via replacing the random denominator by its expectation, i.e. by $\mathbb{E}[f/g] \approx f/\mathbb{E}[g]$, where f is nonrandom [12, 13]. This approximation is exact in the thermodynamic limit, i.e. for large enough t, for example when $t > 20$. One can obtain higher order improvements of the approximation e.g. by a Taylor expansion around the expectation. The quality of such an approximation has been investigated in the context of mean-field theory by Markov Chain Monte Carlo (MCMC) simulations. Based on extensive experimental evidences, for example in [13, 23], the first-order approximation competes with more refined techniques such as the

TAP method [9]. Moreover, for large enough t, as mentioned earlier, the approximation becomes almost exact and the higher order approximation terms diminish. [2]

By replacing $e(t)$ with the value obtained in Equation 3, for large enough t, Equation 4 amounts to

$$\frac{\partial d_v(t)}{\partial t} = \frac{2\eta + m}{2(\eta + m)} \times \frac{d_v(t)}{t} = \frac{\alpha d_v(t)}{t} \tag{5}$$

where $\alpha = \frac{2\eta+m}{2(\eta+m)}$.

To solve Equation 5, we need to find the initial degree of vertex v, i.e. the number of edges v finds when it is added to the network at t_v. At time interval t_v, v finds m edges due to preferential attachment, and it expects to find $\frac{\eta m}{e(t_v)}$ edges due to transitivity. Therefore, its expected initial degree will be $m + \frac{\eta m}{(m+\eta)t_v}$.

Then, using the *continuum theory* [2], we obtain

$$d_v(t) = \left(m + \frac{\eta m}{(m+\eta)t_v}\right)\left(\frac{t}{t_v}\right)^{\alpha} \tag{6}$$

particularly

$$d_v(\mathbf{T}) = \left(m + \frac{\eta m}{(m+\eta)t_v}\right)\left(\frac{\mathbf{T}}{t_v}\right)^{\alpha} \tag{7}$$

If v is added to the network at a large enough time interval (i.e., t_v is larger than a lower bound L), Equations 6 and 7 can be written as

$$d_v(t) = m\left(\frac{t}{t_v}\right)^{\alpha} \tag{8}$$

and

$$d_v(\mathbf{T}) = m\left(\frac{\mathbf{T}}{t_v}\right)^{\alpha} \tag{9}$$

The probability that at time interval $\mathbf{T}$ a vertex v has a degree $d_v(\mathbf{T})$ smaller than k is

$$\Pr[d_v(\mathbf{T}) < k] = \Pr[m\left(\frac{\mathbf{T}}{t_v}\right)^{\alpha} < k] = \Pr[t_v > \frac{\mathbf{T} \times m^{\frac{1}{\alpha}}}{k^{\frac{1}{\alpha}}}] \tag{10}$$

and

$$\Pr[d_v(\mathbf{T}) < k] = 1 - \Pr[t_v \leq \frac{\mathbf{T} \times m^{\frac{1}{\alpha}}}{k^{\frac{1}{\alpha}}}] \tag{11}$$

We suppose that the vertices are added to the network at equal time intervals $\Pr[t_v] = \frac{1}{\mathbf{T}}$. Putting it into Equation 11, we get

$$\Pr[d_v(\mathbf{T}) < k] = 1 - \frac{\mathbf{T} \times m^{\frac{1}{\alpha}}}{\mathbf{T} \times k^{\frac{1}{\alpha}}} = 1 - \left(\frac{m}{k}\right)^{\frac{1}{\alpha}} \tag{12}$$

[2] In our MCMC simulations with $1,000$ runs, the approximation is unbiased, i.e. the difference between the mean of the empirical distribution and the approximated quantity is only 0.061 times the standard deviation.

Then, the degree distribution $\Pr[k]$ can be computed as

$$\Pr[k] = \frac{\partial \Pr[d_v(\mathbf{T}) < k]}{\partial k} = \frac{m^{\frac{1}{\alpha}}}{\alpha} \times k^{-(1+\frac{1}{\alpha})} \tag{13}$$

which means $\Pr[k] \sim k^{-(1+\frac{1}{\alpha})}$. Therefore, we have a power law degree distribution $\Pr[k] \sim k^{-\gamma}$, where

$$\gamma = 1 + \frac{1}{\alpha} = \frac{4\eta + 3m}{2\eta + m} = 2 + \frac{m}{2\eta + m} \tag{14}$$

2.3 The small world property

Reuven Cohen and Shlomo Havlin [8] showed that scale-free networks have a small diameter. In particular, they proved that the scale-free networks with $2 < \gamma < 3$ have a very small diameter which is proportional to $\mathbf{lnln}n$. They also showed that for $\gamma = 3$ the diameter is proportional to $\frac{\mathbf{ln}n}{\mathbf{lnln}n}$, and for $\gamma > 3$ it is proportional to $\mathbf{ln}n$. In all cases the scale-free network satisfies the small-world property. We note that here the diameter is the mean distance between vertices. As Equation 14 indicates, for the η model we have: $2 \leq \gamma \leq 3$. Particularly, for non-zero values of η and m, we have $2 < \gamma < 3$. This means that the η model satisfies the required conditions, i.e. it produces graphs with the small-world property where the diameter is proportional to $\mathbf{ln}n$.

2.4 Clustering coefficient

In this section, we provide a lower bound on the clustering coefficient of the networks produced by the η model, which is independent of the network size and depends only on the η and m parameters.

Watts and Strogatz [28] defined the clustering coefficient of a network as[3]

$$\langle CC \rangle = \frac{1}{n} \sum_{v \in V_{\mathbf{G}}} \langle CC_v \rangle \tag{15}$$

where n is the number of vertices of the network and

$$\langle CC_v \rangle = \frac{C_v}{\binom{d_v}{2}} \tag{16}$$

where C_v is the number of edges among the neighbors of v. $\langle CC_v \rangle$ is called the local clustering coefficient of v.

For a network produced by the η model, C_v can be written as

$$C_v = \sum_{t=t_v}^{\mathbf{T}} \left(\langle C_v \rangle_T (t) + \langle C_v \rangle_P (t)\right) \tag{17}$$

where

[3] An alternative definition of the clustering coefficient which is also widely used, was introduced by Barrat and Weigt [4]: $\frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}}$.

- $\langle C_v \rangle_P (t)$ is the number of edges between neighbors of v which are added to $\mathbf{G}$ during the *preferential attachment* step at time interval t, and
- $\langle C_v \rangle_T$ is the number of edges between neighbors of v which are added to $\mathbf{G}$ during the *transitivity* step at time interval t.

Then, for a vertex v, at every time interval $t \geq t_v$, we define $\tau_v(t)$ as

$$\tau_v(t) = \sum_{t'=t_v}^{t} \langle C_v \rangle_T (t') \tag{18}$$

We have

$$C_v \geq \tau_v(\mathbf{T}) \tag{19}$$

Therefore

$$\langle CC_v \rangle \geq \frac{\tau_v(\mathbf{T})}{\binom{d_v(\mathbf{T})}{2}} \tag{20}$$

Suppose that v is added to the network at a time interval greater than a lower bound L (i.e. $t_v \geq L$) such that we can use Equation 8 to describe its degree. In the following, we compute $\tau_v(\mathbf{T})$.

For $t \geq t_v$, τ_v satisfies the dynamical equation

$$\frac{\partial \tau_v(t)}{\partial t} = \frac{\eta d_v(t)}{2e(t)} = \frac{\eta m t^{\alpha-1}}{2(\eta+m){t_v}^{\alpha}} \tag{21}$$

Then, at time interval $\mathbf{T}$, we will have:

$$\tau_v(\mathbf{T}) - \tau_v(t_v) = \int_{t_v}^{\mathbf{T}} \frac{\eta m}{2(m+\eta){t_v}^{\alpha}} \times t^{\alpha-1} \partial t \tag{22}$$

To solve this dynamical equation, we need to find $\tau_v(t_v)$. Since at time interval t_v vertex v finds $m + \frac{\eta m}{(m+\eta)t_v}$ edges, $\tau_v(t_v)$ will be:

$$\tau_v(t_v) = \frac{\eta \times \left(m + \frac{\eta+m}{(m+\eta)t_v}\right)}{2(m+\eta)t_v} \geq \frac{\eta m}{2(m+\eta)t_v} \tag{23}$$

Therefore after solving the integral of Equation 22, we will have

$$\tau_v(\mathbf{T}) \geq mK \times \left(\frac{\mathbf{T}^{\alpha}}{2{t_v}^{\alpha}} - \frac{1}{2}\right) + \frac{\eta m}{2(m+\eta)t_v} \tag{24}$$

$$\geq \frac{mK\mathbf{T}^{\alpha}}{2{t_v}^{\alpha}} - \frac{mK}{2} \tag{25}$$

where $K = \frac{\eta}{\alpha \times (m+\eta)} = \frac{2\eta}{2\eta+m}$.

Now, we use Equation 20 to find a lower bound for $\langle CC_v \rangle$:

$$\langle CC_v \rangle \geq \frac{\tau_v(\mathbf{T})}{\binom{d_v(\mathbf{T})}{2}} \geq \frac{2\tau_v(\mathbf{T})}{d_v(\mathbf{T})^2} \geq \frac{K{t_v}^{\alpha}}{m\mathbf{T}^{\alpha}} - \frac{K{t_v}^{2\alpha}}{m\mathbf{T}^{2\alpha}} \tag{26}$$

Let v be a vertex such that $L \leq t_v \leq \mathbf{T}$. Up to now, we have computed a lower bound for $\langle CC_v \rangle$. Now, we want to compute a lower bound for the clustering coefficient of the network induced by the vertices added to the network at time intervals $t_L, t_{L+1}, \ldots, t_{\mathbf{T}}$. We refer to this quantity as $\langle CC \rangle$ since it is almost the clustering coefficient of the whole network (compared to $\mathbf{T}$, L is very small and only for few vertices we cannot use Equation 8 to express the degree).

Using Equations 15 and 26, we obtain

$$\langle CC \rangle \geq \frac{1}{\mathbf{T}-L+1} \sum_{t_v=L}^{\mathbf{T}} \left(\frac{K{t_v}^{\alpha}}{m\mathbf{T}^{\alpha}} - \frac{K{t_v}^{2\alpha}}{m\mathbf{T}^{2\alpha}}\right) \tag{27}$$

A simple form of the Riemann sum [26] says

$$\sum_{x=a}^{b} x^r \geq \int_{a-1}^{b} x^r \partial x$$

where $r, a, b > 0$.

This inequality and Equation 27 yield

$$\langle CC \rangle \geq \frac{1}{\mathbf{T}-L+1} \int_{L-1}^{\mathbf{T}} \left(\frac{K{t_v}^{\alpha}}{m\mathbf{T}^{\alpha}} - \frac{K{t_v}^{2\alpha}}{m\mathbf{T}^{2\alpha}}\right) \partial t_v \tag{28}$$

After solving the integral, we obtain

$$\langle CC \rangle \geq \frac{K}{m(\alpha+1)} - \frac{K}{m(2\alpha+1)} \tag{29}$$

$$= \frac{2\eta(\eta+m)}{m(4\eta+3m)(3\eta+2m)} \tag{30}$$

Therefore, a lower bound is provided for the clustering coefficient of a η network, which is independent of the network size and depends on the η and m parameters. We refer to Equation 30 as B.

3 Simulating real-world networks

In this section, we consider several real-world networks, with different specifications and from different domains including biology, technology, social and information networks, and aim to simulate them using the η model. Table 2 summarizes the characteristics of different real-world networks and the networks simulating them. Note that we only describe one way of simulating the real-world networks by the η model which is not unique and the only existing way. In all simulated networks, the initial graph simply consists of two vertices connected by an edge.

Table 2: Real-world networks and the equivalent networks produced by the η model. $\mathcal{C}$ and $\mathcal{C}_\eta$ are clustering coefficient of the real-world networks and clustering coefficient of the networks produced by the η model, respectively. $\mathcal{C}_{BA}$ is the clustering coefficient of the simulated network if transitivity is not used.

Real-world networks				**Simulated networks**				
Network	# vertices	# edges	$\mathcal{C}$	m	η	# edges	$\mathcal{C}_\eta$	$\mathcal{C}_{BA}$
electronic circuits	$24,097$	$53,248$	0.03	2	0.23	$53,121$	0.034	0.009
email address books	$16,881$	$57,029$	0.13	3	0.5	$58,041$	0.11	0.0047
marine food web	135	598	0.23	4	0.54	599	0.24	0.148
neural network	307	$2,359$	0.28	5	2.8	$2,341$	0.29	0.098
Roget's thesaurus	$1,022$	$5,103$	0.15	4	1.4	$5,389$	0.14	0.038

The first real-world network studied here is the *electronic circuits* network. In this network vertices are electronic components e.g., logic gates in digital circuits and resistors, capacitors and diodes in analogic circuits, and edges are the wires [5]. It has $24,097$ vertices and $53,248$ edges and its clustering coefficient is 0.030. In order to simulate this network, we produce an η graph with these parameters: $m = 2$ and $\eta = 0.23$ and it has the same number of vertices as the electronic circuits network. The graph produced by the η model has $53,121$ edges, its clustering coefficient is 0.034 and its degree distribution is depicted in Figure 1(a).

The second real-world network is the network of *email address books* [21]. In this network, vertices represent computer users and an edge is drawn from user A to user B if B's email address appears in A's address book. This network has $16,881$ vertices and $57,029$ edges and its clustering coefficient is 0.13. We simulate this network by the η model using the following parameters: $m = 3$ and $\eta = 0.5$ and the number of vertices in the produced graph is $16,881$. The clustering coefficient of the simulated network is 0.11. However, if we remove transitivity from the network (and produce a BA network), its clustering coefficient will be only 0.0047. Figure 1(b) presents degree distribution of the simulated network.

The next two real-world networks are biological networks. In the *marine food web* network, vertices represent species in an ecosystem and an edge from vertex A to vertex B indicates that A preys on B [14] and [7]. This network has 135 vertices and 598 edges and its clustering coefficient is 0.23. The following parameters are used by the η model to simulate this network: $m = 4$, $\eta = 0.54$, and number of vertices is 135. The produced graph has 599 edges and its clustering coefficient is 0.24. Figure 1(c) presents degree distribution of the networks simulated by the η model.

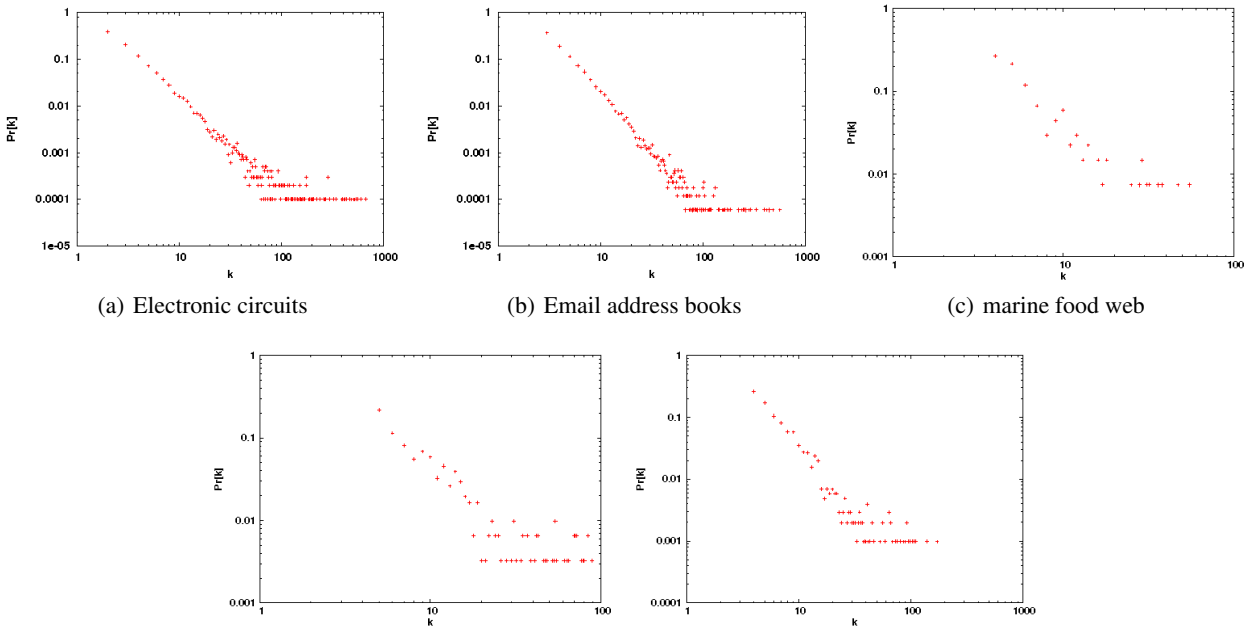
The other important class of biological networks are *neural networks*. The neural network of the nematode C. Elegans reconstructed by White et al. [29] has 307 vertices and $2,359$ edges and its clustering coefficient is 0.28. We simulate it by a η network with $m = 5$ and $\eta = 2.8$. The clustering coefficient of the produced graph is 0.29. Degree distribution of the simulated network is shown in Figure 1(d).

The last real-world network investigated in this paper is the *Roget's thesaurus* network [17]. Each vertex of the graph corresponds to one of the $1,022$ categories in the $1,879$ edition of Peter Mark Roget's Thesaurus of English Words and Phrases. An edge is drawn between two categories if Roget gave a reference to the latter among the words and phrases of the former, or if the two categories were related to each other by their positions in Roget's book. This network has $5,103$ edges and its clustering coefficient is 0.15. We simulate it by a η network with $m = 4$ and $\eta = 1.4$. The produced graph has $5,389$ edges and its clustering coefficient is 0.14. Figure 1(e) presents degree distribution of the simulated network.

Note that when the graph is dense, since m is large, the preferential attachment step has a significant effect on the clustering coefficient. However, the bound B does not consider the clustering coefficient resulted by preferential attachment and as a result, for dense graphs (e.g., the *neural network* graph) it is not tight. In summary, the bound B is always tight for the clustering coefficient resulted by transitivity, and it is tight for the clustering coefficient of a network that is not very dense.

3.1 Empirical evaluation of the η model

In this section, we empirically evaluate the different properties of the η model. In order to investigate the impact of η, we fix m to 2 and n to $10,000$, and produce networks with different η: 0.4, 0.8, 1.2, 1.6, 2. Figure 2 illustrates the degree distributions of the produced networks. If η is set to 0, a BA network is obtained. As we see in the figure, the degree distributions follow a power-law. Furthermore, by increasing η, the exponent γ slowly increases which is consistent with Equation 14. Figure 3(a) compares the clustering coefficient of the networks and the bound B obtained in Equation 30. In the produced networks, m is 2, n is $10,000$ and η varies between 0.4 and 2. It shows that by increasing the clustering coefficient, B increases as well. Table 3 summarizes the characteristics of the simulated networks. In the produced networks, by increasing η, the clustering coefficient, average degree increase and the diameter de-



(a) Electronic circuits (b) Email address books (c) marine food web

(d) nematode C. Elegans. (e) Roget's thesaurus

Figure 1: Degree distributions of the η networks produced for the different real-world networks.

Table 3: Diameter, clustering coefficient, and average degree of networks produced by the η model for different values of η. n is set to $10,000$ and m is set to 2.

η	diameter	clustering coefficient	Avg. degree
0	5.28	0.0045	4
0.4	4.91	0.108	4.701
0.8	4.72	0.171	5.432
1.2	4.21	0.204	6.149
1.6	3.67	0.244	6.900
2	3.25	0.27	7.679

creases.

As depicted in Equations 6, 14 and 30, another parameter affecting the η networks is m. In order to evaluate the influence of m, we fix η to 1 and n to $10,000$, and produce networks with different values for m: 2, 3, 4 and 5. Figure 3 shows degree distributions of the produced networks. As depicted in the figure, the degree distributions follow a power-law. Similar to η, increasing m slightly increases the exponent γ, which is consistent with Equation 14. Figure 3(b) compares the clustering coefficient of the networks and the bound B obtained in Equation 30. In the produced networks, n is $10,000$ and η is 1 and m varies between 2 to 5. It shows that by decreasing B, the clustering coefficient decreases as well and as Equation 30 says, increasing m, reduces B. In Table 4, we describe the specifications of the networks. By increasing m, both the clustering coefficient and the diameter decrease but the average degree increases.

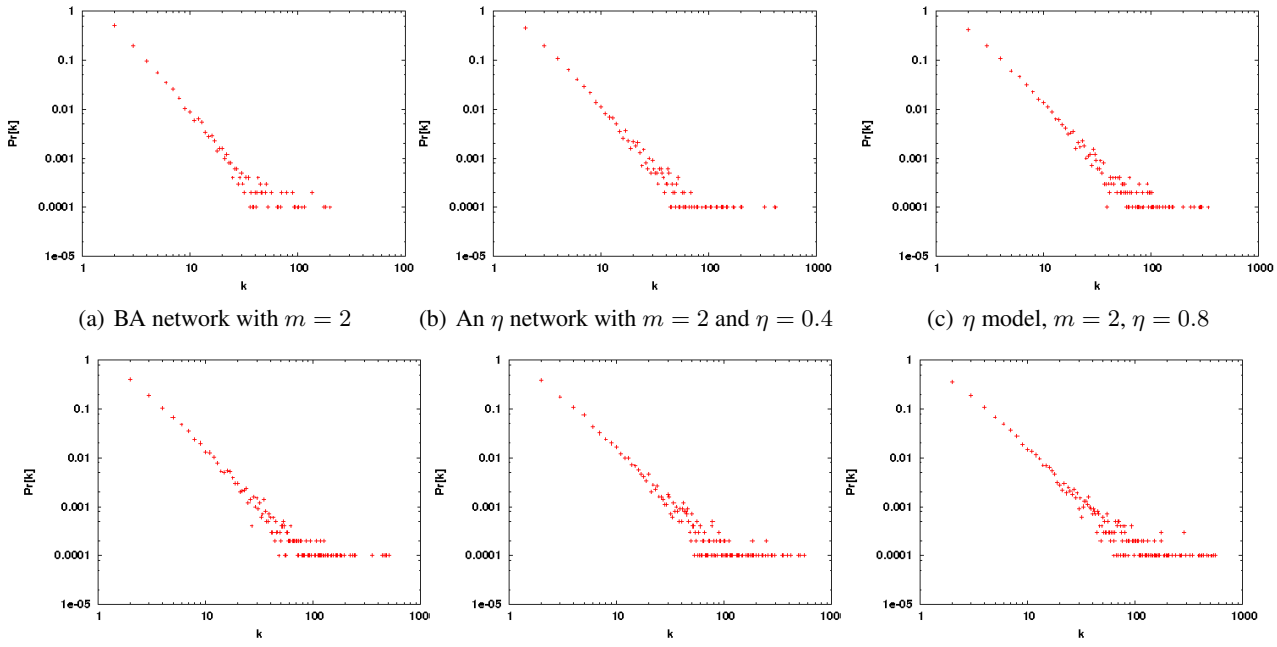
Table 4: Diameter, clustering coefficient, and average degree of networks produced by the η model for different values of m. n is set to 10000 and η is set to 1.

m	diameter	clustering coefficient	Avg. degree
2	4.34	0.19	5.826
3	3.88	0.09	7.804
4	3.32	0.0638	9.91
5	3.47	0.05	11.928

4 Related work

In [1], a power-law model $P(\alpha, \beta)$ is proposed as follows: let y be the number of vertices with degree x. $P(\alpha, \beta)$ assigns uniform probability to all graphs with $y = e^{\alpha}/x^{\beta}$. The authors study the giant component and the evolution of random graphs in this model. The authors of [27] present a model to explain social network searchability. Their model defines a class of searchable networks and a method for searching them.

Chung and Lu [6] consider a family of random graphs with a given expected degree sequence. In this model each edge is selected independently with probability proportional to the product of the expected degrees of its endpoints. Eubank et al. [11] show that many basic characteristics of the social network of the city of Portland, Oregon, USA, are well-modeled by the random graph model of Chung and Lu. They also present approximation algorithms for computing basic structural properties such as clustering coefficients and shortest paths distribution.



(a) BA network with $m = 2$ (b) An η network with $m = 2$ and $\eta = 0.4$ (c) η model, $m = 2$, $\eta = 0.8$

(d) An η network with $m = 2$ and $\eta = 1.2$ (e) η model, $m = 2$, $\eta = 1.6$ (f) η model, $m = 2$, $\eta = 2$

Figure 2: Comparison of degree distributions of a BA network and five η networks having different values of η.

In [20], the authors formulate models of the time evolution of the networks that obtain and lose vertices over time. They show that the model generates networks with power-law degree distributions. In their models new vertices obtain edges by preferential attachment, but the number of added vertices is equal to the number of deleted vertices. In [32] and [33], the authors study and analyze different properties such as degree distribution, clustering coefficient, average path length and phase transition of an evolving email network model.

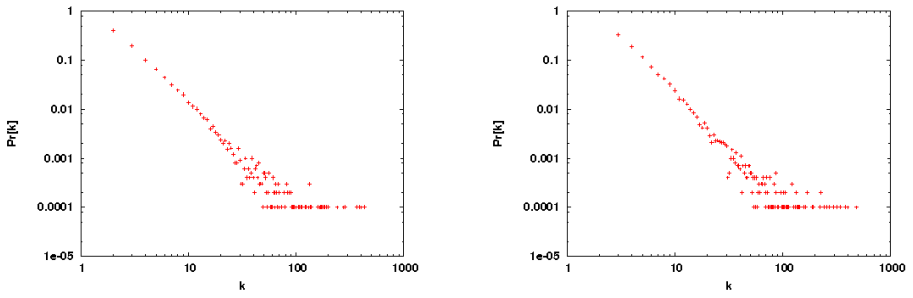
Takemoto and Oosawa [25] propose a model for evolving networks by merging complete graphs (cliques) as building blocks. The model shows power-law degree distribution, power-law clustering spectra and high average clustering coefficients independent of the size of network. However, most real-world networks are formed in a different way: they *grow* over time by obtaining new vertices, rather than by merging cliques.

Serrano, Krioukov and Boguna [24] show that a class of hidden variable models with underlying metric spaces are able to reproduce specific properties (such as topology) in real-world networks. Li and Maini [19] propose an evolving network model that produces community structures. The model is based on two mechanisms: the inner-community preferential attachment and inter-community preferential attachment. However, while their theoretical and numerical simulations show that this network model has community structure, they do not provide a theoretical analysis for the clustering coefficient of the model. Furthermore, their numerical simulations show that the clustering coefficient of their model is still significantly lower than the clustering coefficient of real-world networks.

Yang and Leskovec [30] model the global influence of a vertex on the rate of diffusion through the network. The same authors in [31] investigate several large scale social, collaboration and information networks and find out that the community overlaps are more densely connected than the non-overlapping parts. Kin and Leskovec [16] propose the Multiplicative Attribute Graphs (MAG) model that employs interactions between the vertex attributes and the network structure. In this model, the probability of having an edge between two vertices is proportional to the attribute link formation affinities. The same authors in [15] present a parameter estimation method for the MAG model which is based on variational expectation maximization.

5 Conclusions

In this paper, we proposed a new model, called the η model, for describing transitivity relations in complex networks. We theoretically analyzed the model and calculated a lower bound on the clustering coefficient of the model which is independent of the network size and depends only on the model's parameters (η and m). We proved that the model satisfies important properties such as power-law degree distribution and the small-world property. We also evaluated the model empirically and showed that it can precisely simulate real-world networks from different domains with dif-



(a) $m = 2, \eta = 1$ (b) $m = 3, \eta = 1$

(c) $m = 4, \eta = 1$ (d) $m = 5, \eta = 1$

Figure 4: Comparison of degree distributions of four η networks having different values of m.

ferent specifications.

References

[1] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC, pages 171–180. ACM, 2000.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, 2002.

[3] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] A. Barrat and M. Weigt. On the properties of small-world network models. *EUROP.PHYS.J.B*, 13:547, 2000.

[5] R. F. Cancho, C. Janssen, and R. V. Sole. Topology of technology graphs: Small world patterns in electronic circuits. *Phys. Rev. E*, 64, 2001.

[6] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6:125–145, 2002.

[7] Joel E. Cohen, Frederic Briand, and Charles M. Newman. *Community Food Webs: Data and Theory (Biomathematics)*, volume 20. Springer, 1990.

[8] R. Cohen and S. Havlin. Scale-Free Networks Are Ultrasmall. *Phys. Rev. Lett.*, 90:058701, 2003.

[9] P.W. Anderson D.J. Thouless and R.G. Palmer. A solution to a solvable model of a spin glass. *Philosophical Magazine*, page 35:593, 1977.

[10] P. Erdos and A Renyi. On the evolution of random graphs. pages 17–61, 1960.

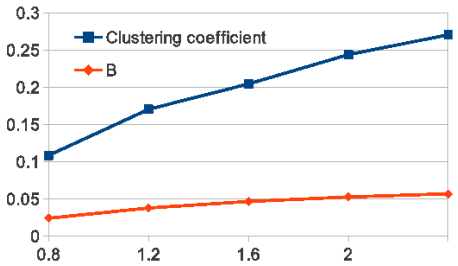
[11] Stephen Eubank, V.S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, and Nan Wang. Structural and algorithmic aspects of massive social networks. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '04, pages 718–727, 2004.

[12] Martin Grötschel, Sven O. Krumke, and Jörg Rambau, editors. *Online Optimization of Large Scale Systems*. Springer, 2001.

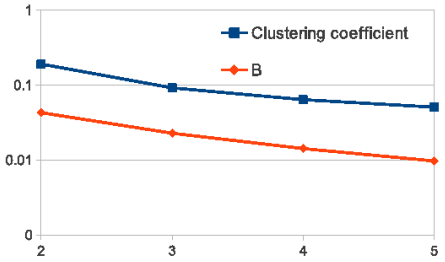
[13] Thomas Hofmann and Joachim M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(1):1–14, 1997.

[14] M. Huxham, S. Beaney, and D. Raffaelli. Do parasites reduce the chances of triangulation in a real food web? *Oikos*, 76:284–300, 1996.

[15] Myunghwan Kim and Jure Leskovec. Modeling social networks with node attributes using the multiplicative attribute graph model. pages 400–409, 2011.



(a) The η parameter.



(b) The m parameter.

Figure 3: Effect of η and m on the clustering coefficient and the bound B.

[16] Myunghwan Kim and Jure Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160, 2012.

[17] Donald E. Knuth. *The Stanford GraphBase - a platform for combinatorial computing.* ACM, 1993.

[18] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.

[19] C. Li and P. K. Maini. An evolving network model with community structure. *Journal of Physics A: Mathematical and General*, 38:9741–9749, 2005.

[20] Cristopher Moore, Gourab Ghoshal, and M. E. J. Newman. Exact solutions for models of evolving networks with addition and deletion of nodes. 74(036121), 2006.

[21] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66(3):035101, 2002.

[22] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[23] Jan Puzicha, Thomas Hofmann, and Joachim M. Buhmann. A theory of proximity based clustering: structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.

[24] M Angeles Serrano, Dimitri Krioukov, and Marian Boguna. Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.*, 100(078701), February 2008.

[25] Kazuhiro Takemoto and Chikoo Oosawa. Evolving networks by merging cliques. *Phys. Rev. E*, 742, 2005.

[26] G.B. Thomas and R.L. Finney. *Calculus and analytic geometry*. Calculus and Analytic Geometry. Addison-Wesley, 1996.

[27] D.J. Watts, P.S. Dodds, and M.E.J. Newman. Identity and search in social networks. *Science*, 296:1302, 2002.

[28] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):409–410, 1998.

[29] J.G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode c. elegans. *Philosophical transactions Royal Society London*, 314:1–340, 1986.

[30] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 599–608. IEEE Computer Society, 2010.

[31] Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '12, pages 1170–1175. IEEE Computer Society, 2012.

[32] Chaopin Zhu and Anthony Kuh. On randomly evolving email networks. In *39th Annual Conference on Information Sciences and Systems*, 2006.

[33] Chaopin Zhu, Anthony Kuh, Juan Wang, and Philippe De Wilde. Analysis of an evolving email network. *Phys. Rev. E*, 74:046–109, 2006.