Modeling Transitivity in Complex Networks

Morteza Haghir Chehreghani, Mostafa Haghir Chehreghani
Xerox Research Centre Europe (XRCE) France, KU Leuven Belgium

Abstract. An important source of high clustering coefficient in real-world networks is transitivity. However, existing approaches for modeling transitivity suffer from at least one of the following problems: i) they produce graphs from 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.

Keywords: real-world graph modeling, clustering coefficient, power-low degree distribution, small-world property.

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) [9], 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, not all models satisfy this condition, for example, the degree distribution of the graphs produced by the ER model converges to a Poisson distribution.

The second property is related to 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 [29]. 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 \( \beta \) model (the Watts-Strogatz model), proposed by Watts and Strogatz [29], 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 $\beta$ model produces an unrealistic degree distribution. The Barabási-Albert (BA) model, proposed by Albert-László Barabási and Réka Albert produces scale-free graphs \cite{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 graphs produced by the BA model is significantly lower than the clustering coefficient of real-world networks. Takemoto and Oosawa \cite{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 networks 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 \cite{17}. The model of \cite{21} 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 \cite{19} is still significantly lower than clustering coefficient of real-world networks. Leskovec et al. \cite{17} 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\cite{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 $\eta$ 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 parameters $\eta$ 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

\footnote{Numerical simulations show that clustering coefficient of a BA network with $n$ vertices is $n^{-0.75}$.}
graphs produced by existing scale-free models, the clustering coefficient of the \( \eta \) 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 \( \eta \) and \( m \). We also empirically investigate 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 the second section, we present the model and theoretically analyze its important properties. In the third section, we empirically evaluate the model and show that it can produce graphs very close to real-world networks. An overview of related work is given in section 4.

2 The \( \eta \) model

In this section, we first present the \( \eta \) model and then we theoretically analyze its important properties like power-law degree distribution, high clustering coefficient and the small-world property.

Algorithm 1 describes the high level pseudo code of the \( \eta \) model proposed for modeling transitivity in complex networks. First a small graph \( G_0 \) is produced. We refer to it as the initial graph. Then, at every time interval \( t \in \{1, \ldots, T\} \), the following steps are performed:

I. growth. A new vertex \( v \) is added to the network \( G \). We denote by \( t_v \) the time of adding \( v \) to \( G \).

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

\[
\frac{d_w(t)}{2e(t)}
\]

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

\[
\frac{\eta d_w(t)}{2e(t)}
\]

where \( \eta \) 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.

The authors of [17] 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 [17] 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.
Algorithm 1 High level pseudo code of the $\eta$ model.

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

1: initialize $G$ by a small graph \\
2: \textbf{for} $t = 1$ to $T$ \textbf{do} \\
3: \{growth\} \\
4: add a new vertex $v$ to $G$ \\
5: \{preferential attachment:\} \\
6: connect $v$ to $m$ existing vertices \{every existing vertex is selected proportional to its degree\} \\
7: \{transitivity:\} \\
8: \textbf{for} every vertex $w \in V_G(t)$ \textbf{do} \\
9: select $w$ with probability $\frac{\eta d_w(t)}{2e(t)}$ \\
10: if $w$ is selected then \\
11: select two neighbors $x$ and $y$ of $w$ uniformly at random \\
12: draw an edge between $x$ and $y$ \\
13: end if \\
14: end for \\
15: end for \\
16: return $G$

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:

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

where $V_G(t)$ denotes vertices of $G$ at time interval $t$. After solving this equation, we obtain

$$e(t) = (m + \eta)t + e(G_0)$$

where $e(G_0)$ denotes the number of edges in the initial graph. For large enough $t$, we sometimes ignore $e(G_0)$ 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 $\eta$ model, vertices (except those added at the very early time intervals) have a power-law degree distribution.
At every time interval \( t \in \{1, \ldots, T\} \), every vertex \( v \) in the network satisfies the following dynamical equation:

\[
\frac{\partial d_v(t)}{\partial t} \approx \sum_{u \in N_v(t)} \left( \frac{\eta d_u(t)}{2e(t)} \times \frac{2}{d_u(t)} \right) + \frac{m d_v(t)}{2e(t)}
\]

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 the 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 \( E[f/g] \approx f/E[g] \), where \( f \) is nonrandom \([11,12]\). 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 \([12,23]\), the first-order approximation competes with more refined techniques such as the TAP method \([27]\). Moreover, for large enough \( t \), as mentioned earlier, the approximation becomes almost exact and the higher order approximation terms diminish.

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

\[
\frac{\partial d_v(t)}{\partial t} = \frac{2 \eta + m}{2(\eta + m)} \times \frac{d_v(t)}{t} = \frac{\alpha d_v(t)}{t}
\]

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

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 initial degree will be \( m + \frac{\eta m}{e(t_v)} \).

Then, using the continuum theory \([2]\), we obtain

\[
d_v(t) = \left( m + \frac{\eta m}{(m + \eta)t_v} \right) \left( \frac{t}{t_v} \right)^\alpha,
\]

particularly

\[
d_v(T) = \left( m + \frac{\eta m}{(m + \eta)t_v} \right) \left( \frac{T}{t_v} \right)^\alpha.
\]
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

$$d_v(t) = m \left( \frac{t}{t_v} \right)^\alpha$$  \hspace{1cm} (8)$$

and

$$d_v(T) = m \left( \frac{T}{t_v} \right)^\alpha$$  \hspace{1cm} (9)$$

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

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

and

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

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

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

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

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

$$\gamma = 1 + \frac{1}{\alpha} = \frac{4\eta + 3m}{2\eta + m} = 2 + \frac{m}{2\eta + m}$$  \hspace{1cm} (14)$$

2.3 The small world property

Reuven Cohen and Shlomo Havlin [8] showed that scale-free networks have a small diameter. Particularly, using analytical arguments, they showed that scale-free networks with $2 < \gamma < 3$ have a very small diameter proportional to $\ln \ln n$. They also showed that for $\gamma = 3$ the diameter is proportional to $\ln \ln n$, and for $\gamma > 3$ it is proportional to $\ln n$. 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 $\eta$ model we have: $2 \leq \gamma \leq 3$. Particularly, for non-zero values of $\eta$ and $m$, we have $2 < \gamma < 3$. This means that the $\eta$ model satisfies the required conditions, i.e. it produces graphs with the small-world property where the diameter is proportional to $\ln n$. 
2.4 Clustering coefficient

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

Watts and Strogatz \[29\] defined the clustering coefficient of a network as

\[
\langle CC \rangle = \frac{1}{n} \sum_{v \in V_G} \langle CC_v \rangle
\]  

(15)

where \( n \) is the number of vertices of the network and

\[
\langle CC_v \rangle = \frac{C_v}{\binom{d_v}{2}}
\]  

(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 \( \eta \) model, \( C_v \) can be written as

\[
C_v = \sum_{t=t_v}^{T} \left( \langle C_v \rangle_T (t) + \langle C_v \rangle_P (t) \right)
\]  

(17)

where

– \( \langle C_v \rangle_P (t) \) is the number of edges between neighbors of \( v \) which are added to \( G \) 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 \( G \) 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

\[
\tau_v (t) = \sum_{t'=t_v}^{t} \langle C_v \rangle_T (t')
\]  

(18)

We have

\[
C_v \geq \tau_v (T)
\]  

(19)

Therefore

\[
\langle CC_v \rangle \geq \frac{\tau_v (T)}{\binom{d_v (T)}{2}}\]

(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 \( \tau_v (T) \).

\footnote{An alternative definition of the clustering coefficient, also widely used, was introduced by Barrat and Weigt \[4\]:

\[ \frac{\text{number of triangles in the network}}{\binom{\text{number of connected triples of vertices}}{3}} \]}
For $t \geq t_v$, $\tau_v$ satisfies the dynamical equation

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

(21)

Then, at time interval $T$, we will have:

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

(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:

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

(23)

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

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

(24)

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

Now, we use Equation 24 to find a lower bound for $\langle CC_v \rangle$:

$$\langle CC_v \rangle \geq \frac{\tau_v(T)}{d_v(T)} \geq \frac{2\tau_v(T)}{d_v(T)^2} \geq \frac{2Kt_v^{\alpha}}{mT^{\alpha}} - \frac{Kt_v^{2\alpha}}{mT^{2\alpha}}$$

(25)

Let $v$ be a vertex such that $L \leq t_v \leq T$. 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 $t_L, t_{L+1}, \ldots, t_T$. We refer to this quantity as $\langle CC \rangle$ since it is almost the clustering coefficient of the whole network (compared to $T$, $L$ is very small and for only a few vertices we can not use Equation 8 to describe the degree).

Using Equations 15 and 25, we obtain

$$\langle CC \rangle \geq \frac{1}{T - L + 1} \sum_{t_v=L}^{T} \left( \frac{Kt_v^{\alpha}}{mT^{\alpha}} - \frac{Kt_v^{2\alpha}}{mT^{2\alpha}} \right)$$

(26)

A simple form of the Riemann sum \[26\] says: $(r, a, b > 0)$

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

This inequality and Equation 26 yield

$$\langle CC \rangle \geq \frac{1}{T - L + 1} \int_{L-1}^{T} \left( \frac{Kt_v^{\alpha}}{mT^{\alpha}} - \frac{Kt_v^{2\alpha}}{mT^{2\alpha}} \right) \partial t_v$$

(27)
After solving the integral, we obtain

\[
\langle CC \rangle \geq \frac{K}{m(\alpha + 1)} - \frac{K}{m(2\alpha + 1)} = \frac{2\eta(\eta + m)}{m(4\eta + 3m)(3\eta + 2m)}.
\]  

(28)

Therefore, a lower bound is provided for the clustering coefficient of a \( \eta \) network, which is independent of the network size and depends on the \( \eta \) and \( m \) parameters. We refer to Equation 28 as \( B \).

3 Empirical simulation

In this section, we empirically evaluate the \( \eta \) model. Our evaluation consists of two parts:

- First, we empirically show that the proposed model can produce graphs very similar to real-world networks. The real-world networks considered here come from different domains. As mentioned earlier, by the \( \eta \) model we aim to develop a network model based on preferential attachment, with the following characteristics: 1) its important properties (clustering coefficient, degree distribution, small-world property) can be verified theoretically, and 2) it precisely simulates real-world networks. To the best of our knowledge, our proposed model is the first model which satisfies the both characteristics. However, we compare the \( \eta \) model with the model of Lescovec et al. [17] which proposes different triangle-closing mechanisms. Their simplest model, called random-random, works remarkably well and has some interesting properties e.g. it is biased towards high-degree vertices. The random-random mechanism is then used in the other parts of the model in [17]. We compare, the \( \eta \) model with this mechanism.

- Second, by performing extensive experiments, we empirically study different properties of the \( \eta \) model and confirm our analytical arguments by numerical simulations.

3.1 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 try to simulate them using the \( \eta \) model and the Lescovec et al. model. Table 1 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 \( \eta \) 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. 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 analog 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 \( \eta \) 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
Table 1. Real-world networks and the equivalent networks produced by the $\eta$ model. $C$, $C_{\eta}$ and $C'$ are clustering coefficient of the real-world network, clustering coefficient of the networks produced by the $\eta$ model, and clustering coefficient of the networks produced by the Lescovec et al. model, respectively. $C_{BA}$ is the clustering coefficient of the simulated network if transitivity is not used.

<table>
<thead>
<tr>
<th>Real-world networks</th>
<th>Simulated networks</th>
<th>The $\eta$ model</th>
<th>The Lescovec et al. model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td># vertices</td>
<td># edges</td>
<td>m $\eta$</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>24,097</td>
<td>53,248</td>
<td>2</td>
</tr>
<tr>
<td>email address books</td>
<td>16,881</td>
<td>57,029</td>
<td>3</td>
</tr>
<tr>
<td>marine food web</td>
<td>135</td>
<td>598</td>
<td>4</td>
</tr>
<tr>
<td>neural network</td>
<td>307</td>
<td>2,359</td>
<td>5</td>
</tr>
<tr>
<td>Roget's thesaurus</td>
<td>1,022</td>
<td>5,103</td>
<td>4</td>
</tr>
</tbody>
</table>

$\eta$ model has 53,121 edges, its clustering coefficient is 0.034 and its degree distribution is depicted in Figure 1(a)\(^3\).

![Degree distribution](image1.png)

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

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

**Fig. 1.** Degree distribution of the $\eta$ networks produced for different real-world networks.

The second real-world network is the network of *email address books* \(^{20}\). 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 $\eta$ model using the following parameters: $m = 3$ and $\eta = 0.5$ and the number of vertices

\(^3\) Simulating this network by the Lescovec et al. model did not finish within a reasonable time (25-30 hours).
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 network simulated by the Lescovec et al. model. The network simulated by the Lescovec et al. model has 45772 edges and its clustering coefficient is 0.37. We note that in the simulated networks, if someone tries to increase the number of edges, the clustering coefficient will increase too and vice versa. Figure 2(a) shows degree distribution of the network simulated by the Lescovec et al. model.

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 [13] and [7]. This network has 135 vertices and 598 edges and its clustering coefficient is 0.23. The following parameters are used by the \( \eta \) 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. Figures 1(c) and 2(b) present respectively degree distribution of the networks simulated by the \( \eta \) model and the Lescovec et al. model. The Lescovec et al. network has 391 edges and its clustering coefficient is 0.376.

The other important class of biological networks are neural networks. The neural network of the nematode C. Elegans reconstructed by White et al. [30] has 307 vertices and 2,359 edges and its clustering coefficient is 0.28. We simulate it by a \( \eta \) network with \( m = 5 \) and 2.8. The clustering coefficient of the produced graph is 0.29. The network produced by the Lescovec et al. model has 837 edges and its clustering coefficient is
Degree distributions of the simulated networks are shown in Figures 1(d) and 2(c), respectively.

The last real-world network investigated in this paper is the Roget’s thesaurus network [16]. 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 \( η = 1.4 \). The produced graph has 5,389 edges and its clustering coefficient is 0.14. The network produced by the Le-scowec et al. model has 2,777 edges and its clustering coefficient is 0.73. Figures 1(e) and 2(d) present degree distribution of the simulated networks.

![Fig. 3. Comparison of degree distributions of a BA network and five η networks having different values of η.](image)

### 3.2 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 \( m = 2 \) and \( n = 10,000 \), and produce networks with different η: 0.4, 0.8, 1.2, 1.6, 2. Figure 3 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 4(a) compares the clustering coefficient of the networks and the bound \( B \) obtained in Equation 28. In the produced networks, \( m = 2, n = 10,000 \) and η varies between 0.4 and 2. It shows that by increasing the clustering coefficient, \( B \) increases as well. Table 2 summarizes
Table 2. Statistics of the \( \eta \) networks for different values of \( \eta \), \( n \) is 10,000 and \( m \) is 2.

<table>
<thead>
<tr>
<th>( \eta )</th>
<th>diameter</th>
<th>clustering coefficient</th>
<th>Avg. degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.28</td>
<td>0.0045</td>
<td>4</td>
</tr>
<tr>
<td>0.4</td>
<td>4.91</td>
<td>0.108</td>
<td>4.701</td>
</tr>
<tr>
<td>0.8</td>
<td>4.72</td>
<td>0.171</td>
<td>5.432</td>
</tr>
<tr>
<td>1.2</td>
<td>4.21</td>
<td>0.204</td>
<td>6.149</td>
</tr>
<tr>
<td>1.6</td>
<td>3.67</td>
<td>0.244</td>
<td>6.900</td>
</tr>
<tr>
<td>2</td>
<td>3.25</td>
<td>0.27</td>
<td>7.679</td>
</tr>
</tbody>
</table>

The characteristics of the produced networks. In the produced networks, by increasing \( \eta \), the clustering coefficient and average degree increase and the diameter decreases.

Fig. 4. Effect of \( \eta \) and \( m \) on the clustering coefficient and the bound \( B \)

As depicted in Equations 6, 14 and 28, another parameter affecting the \( \eta \) networks is \( m \). In order to evaluate the influence of \( m \), we fix \( \eta \) to 1 and \( n \) to 10,000, and produce networks with different values for \( m \): 2, 3, 4 and 5. Figure 5 shows degree distributions of the produced networks. As depicted in the figure, the degree distributions follow a power-law. Similar to \( \eta \), increasing \( m \) slightly increases the exponent \( \gamma \), which is consistent with Equation 14. Figure 4(b) compares the clustering coefficient of the networks and the bound \( B \) obtained in Equation 28. In the produced networks, \( n \) is 10,000 and \( \eta \) is 1 and \( m \) varies between 2 to 5. It shows that by decreasing \( B \), the clustering coefficient decreases as well and as Equation 28 says, increasing \( m \), reduces \( B \). In Table 3, we describe the specifications of the networks. By increasing \( m \), both the clustering coefficient and the diameter decrease but the average degree increases.

Table 3. Statistics of the \( \eta \) networks for different values of \( m \). \( n \) is 10000 and \( \eta \) is 1.

<table>
<thead>
<tr>
<th>( m )</th>
<th>diameter</th>
<th>clustering coefficient</th>
<th>Avg. degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4.34</td>
<td>0.19</td>
<td>5.826</td>
</tr>
<tr>
<td>3</td>
<td>3.88</td>
<td>0.09</td>
<td>7.804</td>
</tr>
<tr>
<td>4</td>
<td>3.32</td>
<td>0.0638</td>
<td>9.91</td>
</tr>
<tr>
<td>5</td>
<td>3.47</td>
<td>0.05</td>
<td>11.928</td>
</tr>
</tbody>
</table>
4 Related work

In [1], a power-law random graph 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 $y = e^{\alpha / x^\beta}$ where self loops are allowed. The authors of [28] present a model to explain social network searchability in terms of characteristics measured along a number of social dimensions. 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. [10] show that many basic characteristics of the social network for the city of Portland, Oregon, USA, are well-modeled by the random graph model of Chung and Lu. They also give fast approximation algorithms for computing basic structural properties such as clustering coefficients and shortest paths distribution.

In [22], the authors formulate models of the time evolution of the networks which acquire and lose vertices during the time. In their models new vertices form links by preferential attachment, but the number of added vertices is equal to the number of deleted vertices. In [33] and [34], the authors study and analyze different properties like 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 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 formed in a different way: they usually grow during the time by obtaining new vertices, rather than by merging complete graphs.

Serrano, Krioukov and Boguna \cite{24} prove that a class of hidden variable models with underlying metric spaces are able to accurately reproduce the self-similarity properties in real-world networks. Li and Maini \cite{19} propose an evolving network model which 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 theoretical analysis for the clustering coefficient of the model.

Yang and Leskovec \cite{31} model the global influence of a vertex on the rate of diffusion through the network. They simulate the number of newly infected vertices as a function of which other vertices got infected in the past. The same authors in \cite{32} examine several large scale social, collaboration and information networks and find that the community overlaps are more densely connected than the non-overlapping parts. This is in contrast to the conventional wisdom that community overlaps are more sparsely connected than the communities themselves. Kin and Leskovec \cite{14} propose the Multiplicative Attribute Graphs (MAG) model which captures interactions between the vertex attributes and the observed network structure. In this model, the probability of having an edge between a pair of vertices depends on the individual attribute link formation affinities. The same authors in \cite{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 $\eta$ model, for describing transitivity relations in complex networks. We theoretically analyzed the model and calculated a lower bound on its clustering coefficient which is independent of the network size and depends only on the model’s parameters ($\eta$ and $m$). We proved that the model satisfies important properties like 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 different specifications.

References

13. 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. M. Kim and J. Leskovec, Modeling Social Networks with Node Attributes using the Multiplicative Attribute Graph Model, UAI, 400-409, 2011.