Extreme Clustering Coefficients in High Edge Density Networks

Yihjia Tsai, Wen-Fa Huang, Chieh-Hsiang Liang, Chen-Han Yao, Kai-Hsiang Tsai
Department of Computer Science and Information Engineering
Tamkang University
151 Ying-chuan Road, Tamsui, Taipei TAIWAN
{eplusplus, huangwenfa,lgsjeffrey ,jj1011}@gmail.com, *hant88@hotmail.com*

Abstract

This paper proposed two models with extreme average clustering coefficients and small path length properties for high edge density network. High density networks are common in the analysis of social networks and biological networks. In addition to those properties, the proposed models indicated that in addition to the existing small-world network model and random network model, there are other network models that may produce clustering coefficients filling the gap between those two models and the maximal achievable clustering coefficients.

1. Introduction

Clustering coefficient [5] is an important vertex measure to estimate the degree of tightness between neighbors of that node. The average clustering coefficient of all the vertices is used to determine whether a graph is a small-world network. A network is small-world if its average clustering coefficient is significantly higher while its average distance between vertices is lower than those of a corresponding random network.

High edge density networks often appear as a clique in a larger network, and are encountered frequently in studies of social networks and biological networks [2]. In studying those high edge density networks, it is quite natural to use existing high clustering coefficient network models such as the popular small-world networks. The small-world network model provides useful insights into the structure and possibly provides explanation to the function and formation of many real-world networks. However, the definition of small-worldness involves the evaluation of a corresponding random network, thus there are some research efforts to define it in a more quantitative way [4]. Some other studies are aiming at designing tunable clustering coefficient algorithms for generating different scale-free networks [1, 3]. Nevertheless, those models are not suitable for high edge density networks. At high edge density, in addition to the trivial clustering coefficient upper bound of 1, and 0 for lower bound, more realistic upper and lower bounds are not available. The two network models proposed in this paper can serve as upper and lower bounds for clustering coefficients. In addition to clustering coefficients, we also show that average network distances have become less important as distinguishing network characteristics in high edge density networks.

2. Definitions

Given a undirected, simple (no self-loops, no multiple edges) network (graph) $G = \{V, E\}$, $V = \{1, 2, \cdots, n\}$ is the set of vertices and $E = \{e_{ij} \mid i, j \in V\}$ is the set of edges. The number of nodes is n and the number of edges is denoted by m. Let $\beta(i)$ be the set of neighboring nodes of i, $\beta(i) = \{v \mid v \in V, e_{vi} \in E\}$, e_{ij} is an edge connecting node i and j, and the number of elements in set $\beta(i)$ is denoted by $\beta_i = |\beta(i)|$ and is equal to its node degree $d_i = \beta_i$. The intersection of two sets $\beta(i)$ and $\beta(j)$ is represented by $\delta(i, j) = \beta(i) \cap \beta(j)$ and the number of elements in that set is $\delta_{i,j} = |\delta(i, j)|$. The clustering coefficient of node i is defined as the ratio of the number of links between neighboring nodes of i and the maximal possible number of links between all of its neighboring nodes, $cc_i = \sum_{\forall j \in \beta(i)} \delta_{i,j} \big/ \beta_i(\beta_i - 1)$. The clustering coefficient of network G is therefore the average of all clustering coefficient of the vertices in the network $cc_G = \sum_{i=1}^{n} cc_i \big/ n$.

978-1-4244-5228-6/09/$26.00 c2009 IEEE

As clustering coefficient of a node i is a value between 0 and 1, thus in order to facilitate the process of presentation, we use filled circle for clustering coefficient of 1 while gray circle represents intermediate values (including zero) of clustering coefficient. We are interested in studying networks with high edge density, to be more specific, networks with more than $(n^2-5n+12)/2$ edges for a size n network. When we divide edge count by the maximum number of possible edges for a size n network, the edge density is $1-4(n-3)/(n^2-n)$, which is close to one for even a moderate size n. At such high edge density, we show some properties related to network clustering coefficients. First, we define a quantity $m^- = (n^2-n)/2-m$ as a measure of the number of edges needed to form a complete network.

Proposition 1 *When* $m>(n^2-5n+12)/2$, *and* $m<(n^2-n)/2$, *the number of gray nodes in a size n network is greater than 1.*

The proof is based on the following inference steps:

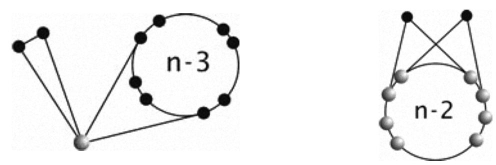
1. A fully connected network is the only connected network without any gray node. When $m<(n^2-n)/2$, at least a gray node will present in the network.
2. When there is exactly one gray node in the network, all other black nodes have a direct link connected to the gray node.
3. The minimum value of m^- for a one-gray node network is $2(n-3)$, which corresponds to $m=(n^2-5n+12)/2$ network.
4. Therefore, for any m^- less than $2(n-3)$, the network must have more than one gray node, thus completes the proof.

The proof of 3. is by dividing $(n-1)$ into k sets, and denoting them as $x_1, x_2, \cdots, x_k$, $(n-1)=\sum_{i=1}^{k} x_i$, and $2 \le x_i \le (n-1)$ for $i=1,2,\cdots,k$. We know that $\left(\sum_{i=1}^{k} x_i\right)^2 = \sum_{i=1}^{k} x_i^2 + 2\sum_{i<j} x_i x_j$, and $\sum_{i<j} x_i x_j = m^-$, therefore, $2m^- = (n-1)^2 - \sum_{i=1}^{k} x_i^2$. The minimum value of m^- occurs when $k=2$ and $x_1=2$, $x_2=n-3$, which corresponds to $\min(m^-) = 2(n-3)$.

3. Network Models

3.1. Maximal Clustering Coefficients

The proposed network model for high clustering coefficient is by adding one edge at a time. At each step, one edge is added and the network is reorganized to produce the highest possible clustering coefficients. For a given n, the process begins with a network that contains exactly $m = (n^2-5n+12)/2$ edges. This network, as shown in Figure 1(a), has a $m^- = 2(n-3)$ and it has only one gray node.

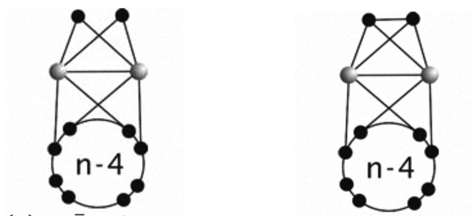


(a) $m^- = 2(n-3)$ **network** **(b)** $m^- = 1$ **network**

Figure 1

The network in Figure 1(b) has a m^- value of 1 and its total number of edges is $m = (n^2-n)/2-1$, it has exactly $(n-2)$ gray nodes. The edges difference between networks of Figure 1(a) and Figure 1(b) is $(2n-7)$, and the difference between the number of gray nodes is $(n-3)$. Thus the minimum increase of gray nodes per edge addition is $2-1/(n-3)$, which is roughly 2 for large n. Suppose the clustering coefficient of the only gray node in Figure 1(a) is denoted as cc_1, $cc_1 = (n-5)/(n-1)$. The network average clustering coefficient is $cc_{net} = 1-(1-cc_1)/n = 1-4/(n(n-1))$.

Now we illustrate the process of forming a network of size n and given m edges, $(n^2-5n+12)/2<m<(n^2-n)/2-1$, by adding one edge at a time and readjusting the resulting network. Adding one edge to the network in Figure 1(a) will produce a network with two less than 1 clustering coefficient nodes, as indicated by the gray node in Figure 2(a). This is achieved by joining one black node on the left wing of the network in Figure 1(a) and one of the black nodes in the right sub-network. Next, breaking apart the edge between the two black nodes on the left wing and joining the other black node to the one on the right sub-network. Now, all nodes in the network other than the two gray nodes, colored in black, have a clustering coefficient of 1. The node degree of one of the two gray nodes in Figure 2(a) is $n-1$ while the total number of edges between these $(n-1)$ neighbors is $(n-3)(n-4)/2+2=(n^2-7n+16)/2$. Thus the clustering coefficient of the gray node is $cc_2=(n^2-7n+16)/(n-1)(n-2)$.

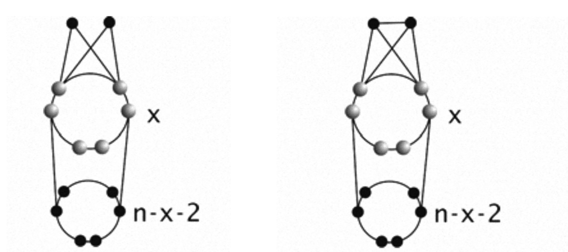


(a) $m^-=2n-7$ **(b)** $m^-=2n-8$

Figure 2

The addition of one edge to network in Figure 2(a) between the two black nodes will produce a network topology, as shown in Figure 2(b), with very similar clustering characteristics as the network in Figure 2(a). All nodes in the network other than the two gray nodes have a clustering coefficient of 1. The node degree of the gray node in Figure 2(b) is $n-1$ while the total number of edges between these $(n-1)$ neighbors is $(n^2-7n+18)/2$. Consequently, the clustering coefficient of the gray node is $cc_2=(n^2-7n+18)/(n-1)(n-2)$. The network average clustering coefficient is $cc_{net}=1-2(1-cc_2)/n$.

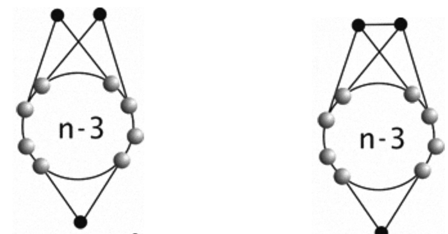
By adding one network edge at a time, starting from network of Figure 1(a), at step k, $k=1,2,\cdots,2(n-3)$, the number of gray nodes in the network is $x=\lfloor (k+1)/2 \rfloor+1$.



(a) $m^-=2(n-3)-k$ **(b)** $m^-=2(n-3)-(k+1)$

Figure 3

As shown in Figure 3(a), when k is odd, the two black nodes on the upper portion of the network are separated. When $k+1$ is even, as indicated by Figure 3(b), the two black nodes on the upper portion of the network is connected. When the number of added edges reaches $k=2n-9$, which is an odd number, the resulting network is shown in Figure 4(a). Consequently, adding one edge to the network in Figure 4(a) will produce the one in Figure 4(b). And two more edges connecting the one black node on the lower part of the network with the two black nodes on the upper part will produce a complete network. If only one edge instead of two is added to the network of Figure 4(b), we will reach the network in Figure 1(b).



(a) $m^-=3$ **(b)** $m^-=2$

Figure 4

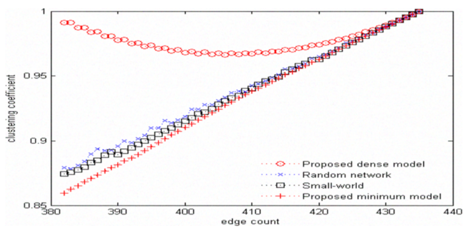
The preceding process can produce a size *n* network with any given *m* edges, $(n^2-5n+12)/2<m<(n^2-n)/2-1$, starting from a network of $(n^2-5n+12)/2$ edges. The increase of gray nodes per edge addition is kept at a minimum value of 2 during the process. Therefore, we can conclude that the proposed network model has reached a maximum network clustering coefficient achievable for that given number of edges.

3.2. Minimal Clustering Coefficients

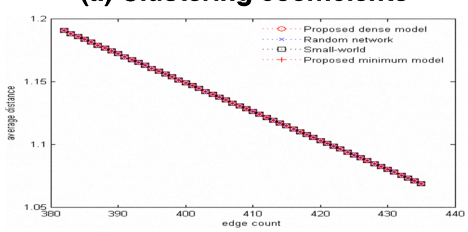
The clustering coefficient for a complete bipartite network is zero and the maximal number of edges for a *n*-node bipartite network is $\lfloor n^2/4 \rfloor$. Thus, for a connected network with edge counts less than $\lfloor n^2/4 \rfloor$ and larger than $(n-1)$, its minimum clustering coefficients is trivially zero. For edge counts larger than $\lfloor n^2/4 \rfloor$, the exact bound for minimal clustering coefficient is unknown. The proposed network model begins with a maximal edge complete bi-partite network and adding edges in a recursive fashion. When the required network edges *m* is more than $\lfloor n^2/4 \rfloor$, define an amount $m^+=m-\lfloor n^2/4 \rfloor$ and connecting nodes within the half sub-network in a bi-partite network fashion. If more edges is needed, then, divide the two sub-sub-network into further smaller networks and connecting edges within those smaller networks according to bi-partite network rules. In the extreme, this can produce a complete network if the number of required network edges is exactly $(n^2-n)/2$.

4. Experiments and Results

To understand the behavior of the proposed network models, we compare the network clustering coefficients and average network distances with two well-known network models: the random network model and the small-world network models. The results for network of size n=30 are shown in Figure 5(a) and 5(b).



(a) Clustering coefficients

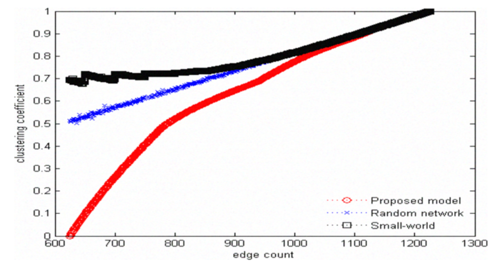


(b) Average network distances

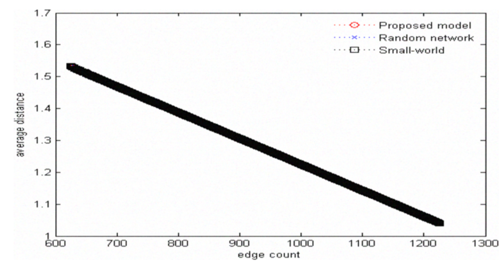
Figure 5 n = 30 networks

For high edge density networks, Figure 5(a), the clustering coefficients of random network are shown in the middle and it increases linearly as the number of edges in the network increases. As expected, the clustering coefficients of the proposed maximum/minimum model produced the highest/lowest curves among the four models. The small-world network model has a higher clustering coefficient than random network. Clustering coefficients of all these models converge to 1 when the number of edges increased to near fully connected network. There is a huge gap between the proposed maximum model and the other three network models. This is a clear indication that there exist other network models which will generate those intermediate clustering coefficients.

Similar results can be observed for a different sized network, as shown in Figure 6(a), which are simulation results for size n=50 networks for a edge count range starting from $\lfloor n^2/4 \rfloor$. This is to show the behavior of the proposed minimum clustering coefficient model in those ranges. As for average network distances, it is rather interesting to observe that all network models have the same average network distances in high edge density.



(a) Clustering coefficients



(b) Average network distances

Figure 6 n = 50 networks

5. Conclusion

In this paper we propose two network models with extreme properties for high edge density networks, that is, for a size n network, the number of edges is m , and $(n^2 - 5n + 12)/2 \leq m \leq (n^2 - n)/2$. This work provide as a foundation for further study of small-world property and the effects of clustering coefficient on various network collective phenomenon. The results also indicate that there exist network models that can produce intermediate clustering coefficients other than small-world network models.

6. References

[1] Fu, P. and K. Liao, "An Evolving Scale-free Network with Large Clustering Coefficient," in ICARCV 2006,

[2] Girvan, M. and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences of the USA, vol. 99, no. 12, pp. 7821-7826, 2002.

[3] Holme, P. and B. J. Kim, "Growing scale-free networks with tunable clustering," Physical Review E, vol. 65, p. 026107, 2002.

[4] Humphries, M.D. and K. Gurney, “Network ‘Small-World-Ness’: A Quantitative Method for Determining Canonical Network Equivalence,” PLoS ONE, vol. 3, no. 4, pp. e0002051, 2008.

[5] Watts, D. and S. Strogatz. “Collective dynamics of small-world networks,” Nature 393, pp. 440-442, 1998.