Egyptian Informatics Journal (2021) 22, 389–400

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

FULL-LENGTH ARTICLE

A two-hop neighbor preference-based random network graph model with high clustering coefficient for modeling real-world complex networks

Natarajan Meghanathan *, Aniekan Essien, Raven Lawrence

Jackson State University, Jackson, MS 39217, USA

Received 15 November 2015; revised 22 April 2016; accepted 19 June 2016
Available online 8 September 2016

KEYWORDS

Two-hop neighbors;
Random network graphs;
Erdos-Renyi model;
Complex network analysis;
Clustering coefficient

Abstract The Erdos-Renyi (ER) random network model generates graphs under the assumption that there could exist a link *u*-*v* between two nodes *u* and *v* irrespective of whether or not the two nodes had a common neighbor before the establishment of the link. As a result, random network graphs generated under the ER model are characteristic of having a low clustering coefficient (a measure of the probability for a link to exist between any two neighbors of a node) and low variation in the node degrees, and hence could not match closely to graphs abstracting real-world networks. In this paper, we propose a random network graph model that gives preference to closing the triangle involving three nodes *u*, *w* and *v* with existing links *u*-*w* and *w*-*v* (i.e., node *v* is strictly a two-hop neighbor of node *u*). Accordingly, when node *u* is looking for a new link to be setup with some other node *x*, we consider *x* along with the two-hop neighbors of *u* and choose one among these nodes with a probability p_{link} as the new neighbor of node *u*. The proposed Two-Hop Neighbor Preference (THNP)-based model generates random graphs whose clustering coefficient decreases with increase in node degree: matching closely to several real-world network graphs that are commonly studied for complex network analysis.

1. Introduction

With the growth of social networks and the availability of data about several real-world networks (such as biological networks and citation networks), there has been a significant interest to analyze these complex networks from a graph theoretic point of view. A crucial step in analyzing a complex real-world network is to extract the degree distribution of the vertices in the network and compare it with well-known distributions (e.g., Poisson, Gaussian, Exponential, Power-law, etc.) of networks [1] generated from theoretical models using the same parameters (such as the average node degree and standard deviation of node degree) as that of the real-world network. Once the

* Corresponding author.
E-mail address: natarajan.meghanathan@jsums.edu (N. Meghanathan).
Peer review under responsibility of Faculty of Computers and Information, Cairo University.

http://dx.doi.org/10.1016/j.eij.2016.06.008
1110-8665 © 2021 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Computers and Artificial Intelligence, Cairo University.

equivalent theoretical graph model for a real-world network is identified, one can effectively analyze the various characteristics of the real-world network based on the parameters and analytic metrics of the graph model identified.

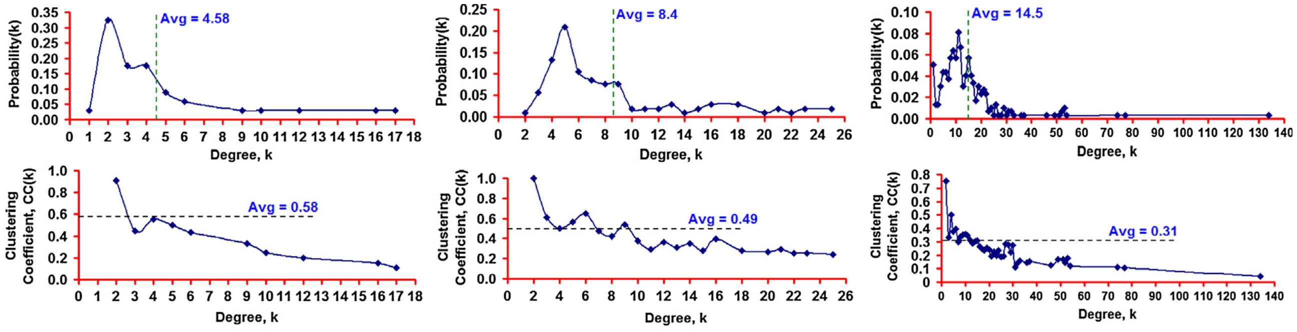
The results of the analysis of certain real-world networks indicate that the degree distribution of the vertices in these networks does not closely match with any of the well-known distributions. For example, we observe (from Fig. 1) that the degree distribution of some sample real-world networks studied in this paper is a Poisson-curve (characteristic of random networks) [2], but with a long tail (characteristic of scale-free networks) or sometimes even with a long head. If these real-world networks are modeled as a random network, then one cannot capture the presence of a few, but appreciable number of vertices that have a significantly larger degree than the rest. Also, a random network cannot capture the decreasing nature of the clustering coefficient of the vertices with increase in node degree (the clustering coefficient of the vertices in a random network is independent of node degree and is simply equal to the probability of a link between any two nodes) [3]. On the other hand, if these real-world networks are modeled as a scale-free network [4], then one would wrongly capture the degree distribution of vertices with lower degree; while scale-free networks are expected to have a larger number of vertices with lower degree, the real-world networks shown in Fig. 1 have only fewer vertices with lower degree. Thus, the motivation of the observations from Fig. 1 is that we need a random network graph model that captures the presence of smaller, but appreciable percentage of vertices with both lower degree and higher degree (i.e., larger variation in node degree). At the same time, we also want the model to capture the complex inverse relationship between clustering coefficient and node degree.

In this paper, we propose that whenever we are looking for a link to be set up for a node *u*, we give preference to the two-hop neighbors of node *u*, rather than to an arbitrary node that may or may not be a two-hop neighbor of node *u* before the establishment of the link. We hypothesize that the proposed approach could lead to the closure of several two-link chains as triangles – contributing to a larger clustering coefficient for the vertices as well as contribute to a larger variation in node degree among the vertices, compared to the well-known Erdos-Renyi (ER) model [5] for random networks. The ER model has a characteristic of generating random network graphs with lower variation in node degree (the degree of a majority of the vertices lies close to the average degree) and lower clustering coefficient (the clustering coefficient for any vertex in an ER random graph is close to the probability of a link between any two nodes).

According to the proposed two-hop neighborhood preference (THNP) approach, if we are to set up a new link for a node *u* and decide whether the node is to be paired with a randomly chosen candidate node *x* (with which node *u* does not yet have a link), we consider node *x* along with the two-hop neighbors of node *u* (i.e., there exists two-hop chains, say *u*-*w*-*v* involving links *u*-*w* and *w*-*v*) and randomly choose one among the nodes (from the expanded candidate list) with a probability (p_{link}) as the neighbor node. The p_{link} value used in the THNP model is the same as the p_{link} value used in the ER model and is obtained by dividing the average node degree of the real-world network by number of nodes – 1.

Unlike the ER-model based random network graphs, the THNP-model based random network graphs exhibit an inverse relationship for the clustering coefficient vs. node degree as well as a larger spectral radius ratio for node degree (i.e., larger variation in node degree) [6]. We evaluate the relative proximity (similarity) of the graphs generated according to the THNP model and the ER model with that of the real-world network graphs by computing the rooted sum of the mean square difference values for the average node degree, average clustering coefficient, average path length and the spectral radius ratio for node degree. We observe the THNP-model based random network graphs to be relatively more similar to the real-world network graphs considered, compared to that of the ER-model based random network graphs.

The rest of the paper is organized as follows: Section 2 discusses related work on theoretical models for generating networks with clustering coefficient values and distribution similar to those observed for real-world networks. Section 3 presents the proposed THNP model for generating random network graphs with a larger clustering coefficient and higher variation in node degree. Section 4 reviews the ER-model, presents the real-world network graphs considered in this paper and illustrates a comparative analysis of the ER-model and the THNP-model with that of the real-world network graphs with respect to the analytic metrics mentioned above. Section 5



Karate Club Network
(34 nodes, 78 edges)

US Politics Book Network
(105 nodes, 441 edges)

C. Elegans Neural Network
(297 nodes, 2148 edges)

Figure 1 Degree distribution and clustering coefficient vs. degree for real-world networks.

concludes the paper. Throughout the paper, we use the terms 'node' and 'vertex', 'link' and 'edge', as well as 'network' and 'graph' interchangeably. They mean the same.

2. Related work

In this section, we present related work in the literature on theoretical models proposed to generate networks whose clustering coefficient distribution resembles close to that for real-world networks and also highlights the unique characteristic (s) that differentiate our proposed THNP model from these existing models. In [14], the authors propose a model for scale-free random graphs (the network size increases with the addition of one node at a time) in which each new vertex is added to m of the existing nodes with a probability proportional to the degree of the existing nodes plus a strictly positive constant (the latter is associated with introducing randomness in node selection). It was observed that the expected clustering coefficient of the scale-free random network generated with the above approach is asymptotically proportional to $\log n/n$. The THNP model differs from the above model because the candidate list of nodes available for setting up a link for a node u is the two-hop neighbors of u and a randomly chosen node v that is originally considered for the link. Among the two-hop neighbors, the nodes with a larger number of common neighbors are given relatively higher preference. Not all the nodes in the network are considered for setting up a new link for node u. This is similar to what happens in a social network when a node/user wishes to set up a new link; the user is more likely to search for friends of his/her friends rather than searching out for some random user in the network. Moreover, if we apply the above $\log n/n$ formulation to the real-world network graphs considered in this paper, the expected clustering coefficients turn out to be very small values (like 0.0076 for the 332-node US Airports Network and 0.045 for the 34-node Karate Club Network) and are independent of the node degree.

In [15], the authors propose a model for creating random intersection graph as follows: Let W be the set of all n vertices; let $D_1, D_2, \ldots, D_n$ be randomly created subsets of the n vertices; two vertices u and v are considered to be linked in the graph if they are in at least s of the n subsets of vertices, where $s \geqslant 1$ is a model parameter. It is observed in [16] that the random intersection graphs exhibit a linear inverse relationship between clustering coefficient and node degree (i.e., the clustering coefficient of a node is proportional to k^{-1} where k is the node degree). In this paper, we observe that the clustering coefficient-degree relationship need not be simply linear; it could be more complicated as is observed in Fig. 12(a)–(f). Hence, the random intersection graph model is not sufficient to model the complex inverse relationship between clustering coefficient and node degree, as seen in several real-world networks.

In [17], the authors proposed a model to generate random graphs with degree distribution that follows power-law [4]. The idea is as follows: (i) Determine the degree distribution of the vertices in a real-world network. (ii) Make a list of the vertices by generating k_v copies of each vertex v, where k_v is the degree of vertex v. (iii) Set up two columns of the list of vertices by doing two different random shuffles of the list of vertices created in step (ii). (iv) Connect the vertex pairs in each row of the two columns. The problem with the above approach is that for real-world networks with several vertices, it could lead to graphs with self-loops and redundant edges. The authors in [17] propose certain fixes to overcome these weaknesses. Nevertheless, the clustering coefficient of the vertices in such power-law based random networks has been observed to be simply dependent on the power-law exponent and the number of nodes, but not on the node degree. Thus, the above model for generating power-law based random networks does not capture the inverse relationship between clustering coefficient and node degree, as observed in several real-world networks, and well captured by the proposed THNP model (see Section 4).

Instead of using the degree sequence alone, there has been some active research (e.g., [18–19]) on generating random graphs based on the number of triangles a node is part of. Such an approach has been observed to yield tunable clustering coefficient. The idea is to input the triangle degree sequence (i.e., the number of triangles each node is part of) and the single-edge degree sequence (i.e., the number of edges incident on a node and not part of any triangles) to generate random graphs by selecting three vertices at a time and creating a triangle between them (if their residual triangle degree is greater than 0) and selecting two vertices at a time and linking them (if their residual single-edge degree is greater than 0). The residual triangle degree and residual single-edge degree of a vertex are respectively the remaining number of triangles and single-edges (compared to the input values) that the vertex is yet to be part of. The above procedure is repeated until the residual triangle degree of the vertices becomes 0 and the residual single-edge degree of the vertices becomes 0. The above idea was first proposed in [18], but there was no clarification on which three vertices for triangle formation and which two vertices for single-edge formation are to be selected in an iteration.

In [19], the authors propose several strategies (e.g., based on the ratio of the triangle degree and the sum of the triangle degrees, ratio of the single-edge degree and the sum of the single-edge degrees, etc.) for selecting vertices for triangle and single-edge formation. A key weakness behind the above approaches is that it would take a significant amount of time to compute the number of triangles and the number of single-edges each node is part of. Also, the convergence time for completing the formation of all triangles will be high (as any three nodes chosen need not have a triangle connecting them). On the other hand, the advantage with our proposed THNP algorithm is that it only needs the estimated value of the probability of a link (p_{link}) between any two nodes in the real-world network as input ($p_{link} = \langle k \rangle / N-1$, where - $\langle k \rangle = 2^{*} L/N$ is the average degree, N and L are the number of nodes and links respectively) and the p_{link} value can be simply computed based on the number of nodes and links in the real-world network graph.

3. Two-Hop Neighbor Preference (THNP) Model

Let there be a total of N nodes among which we need to generate a random network with a larger clustering coefficient and higher variation in node degree. Let p_{link} be the probability for a link between any node pair considered. Initially, all nodes are considered to be isolated in the network. We consider a set S of all possible node pairs (a network of N nodes has $N(N-1)/2$ node pairs). The network generation proceeds in iteration. In each iteration, we pick a node pair u-x from the set S and randomly pick one among the two vertices (u or x) as the node for which we attempt to set up a link during that iteration. For the sake of discussion, we assume (without loss of generality that between the two vertices u or x) that vertex u gets selected as the vertex for which we attempt to set up a link. Let C be the list of candidate vertices with which we could set up a link for u. We add vertex x as the first vertex to the list C. We then consider all the two-hop neighbors v of u (i.e., connected through a chain u-w-v where w is the direct neighbor of u) that are not directly connected to vertex u and add them to the list C. We randomly pick a vertex (say, vertex v) from the list C. We generate a random number *randNum* in the range [0...1]. If $randNum \leqslant p_{link}$, we set up a link between u and v; otherwise not. If a link between u and v gets set up, we remove the pair u-v from the set S. Irrespective of the outcome of the iteration, we remove the pair u-x from the set S and continue with the next iteration. We repeat the above steps in each iteration until the set S becomes empty. Fig. 2 presents the pseudo code for the THNP-algorithm.

With regard to the time-complexity, lines 1–5 are executed in $\Theta(N^2)$ time, where N is the number of nodes in the network. The while loop from lines 6–32 runs for a total of $N(N-1)/2 = \Theta(N^2)$ iterations – one iteration for each node pair. For each such iteration, lines 7–17 and 25–31 can be executed in $\Theta(1)$ time; lines 18–24 involve a traversal of the two-hop neighborhood of a vertex. If $K = 2L/N$ is the average degree of a node [3] in the final network with a total of L links, each execution of lines 18–24 takes on average $\Theta(K^2)$ time (as essentially the neighborhood of each of the K neighbors of a node is explored as part of the search for two-hop neighbors). Hence, lines 18–24 could be considered to be of complexity $\Theta(L^2/N^2)$ for each of the $\Theta(N^2)$ iterations. Thus, each iteration of the while loop from lines 7–30 could be considered to take $\Theta(L^2/N^2)$ time and for a total of $\Theta(N^2)$ iterations, the time-complexity for lines 6–32 is $\Theta(L^2)$. Hence, the overall time-complexity of the algorithm to generate a random graph under the THNP model is $\Theta(N^2)$ – for lines 1–5 + $\Theta(L^2)$ – for lines 6–32 = $\Theta(N^2 + L^2)$.

3.1. Relevance to social networks

As new links get added, the number of two-hop neighbors of a node increases and there are high chances that a node u gets linked with one of its two-hop neighbors rather than to a vertex not in its two-hop neighborhood. In social networks, if a user intends to pair with someone in the network, the user is more likely to prefer pairing with someone who is known to

```
--------------------------------------------------------------------------------
Input: Number of Nodes, $N$; Probability of a Link, $p_{link}$
Output: Set of Links, $E$
Auxiliary Variables: Set of Node Pairs, $S$; Candidate List of Vertices, $C$; Candidate Vertex
Initialization: $E = \phi$
Begin THNP Algorithm
1     for every vertex $u = 1$ to $N$ do
2         for every vertex $v = u+1$ to $N$ do
3             $S = S \cup \{(u, v)\}$
4         end for
5     end for
6     while ($S \neq \phi$) do
7         Select a pair $(u, x)$ randomly from $S$
8         $S = S - \{(u, x)\}$
9         $C = \phi$
10        Generate a random number $r$ in the range [0...1]
11        if $r \leq 0.5$ then
12            $C = C \cup \{x\}$
13            Candidate Vertex = $u$
14        else
15            $C = C \cup \{u\}$
16            Candidate Vertex = $x$
17        end if
18        for every vertex $w \in$ Neighbor(Candidate Vertex) do
19            for every vertex $v \in$ Neighbor($w$) do
20                if $v \notin$ Neighbor(Candidate Vertex) then
21                    $C = C \cup \{v\}$
22                end if
23            end for
24        end for
25        Pick a vertex $v$ randomly from the list $C$
26        $C = C - \{v\}$
27        Generate a random number randNum in the range [0...1]
28        if $randNum \leq p_{link}$ then
29            $E = E \cup \{(Candidate\ Vertex - v)\}$
30            $S = S - \{(Candidate\ Vertex, v)\}$
31        end if
32    end While
33    return $E$
End THNP Algorithm
--------------------------------------------------------------------------------
```

Figure 2 Pseudo code for the two-hop neighbor preference-based random network generation model.

his/her neighbors rather than pairing with an unknown stranger. Several social media networks, including Facebook, have been designed to suggest friends of friends as possible candidates to send friend requests. Thus, the Two-Hop Neighbor Preference model is very well applicable to model friendship-based interactions in social networks. Note that if more neighbors of the *Candidate Vertex* have a particular vertex as their two-hop neighbor, then the vertex has greater chances of becoming a neighbor of the *Candidate Vertex*. This is in tandem with how links are formed in social networks like Facebook. If a user *A* gets a Friend Request from two users *B* and *C* and is in a situation of accepting just one of these Friend Requests, then user *A* is more likely to accept the Friend Request from the user who has relatively larger number of common friends.

3.2. Example to illustrate the execution of the THNP model

Fig. 3 presents an example to illustrate the execution of an iteration of the THNP model to set up a link. Let a pair (3, 5) be removed from the set *S* and considered for the possibility of setting up a link for *Candidate Vertex* = 3 (chosen randomly among vertices 3 and 5). Since vertex 3 is chosen as the *Candidate Vertex*, the other vertex (vertex 5) is included to the list *C* of vertices that could be paired with the *Candidate Vertex*. We now look for the two-hop neighbors of the *Candidate Vertex* (vertex 3) and include them to the list *C*. Note that a two-hop neighbor vertex could be included multiple times to the list *C*, with one entry per neighbor. We observe vertex 8 having two entries in the list *C* because of it being the neighbor of vertices 4 and 7. The other two-hop neighbors (vertices 1 and 6) have only one entry in the list *C* as they are the neighbor of only one of the neighbors of the *Candidate Vertex* 3. We randomly pick a vertex from the list *C* = {1, 8, 5, 6, 8}. As is observed in the simulations, vertices that have multiple entries in the list *C* have a relatively larger chance of being selected (increasing the number of closed triangles and hence increasing the average clustering coefficient of the entire network): we observe vertex 8 to be selected as the vertex with which the *Candidate Vertex* 3 will set up a link.

3.3. Impact of the preference for more commonly occurring two-hop neighbors

Fig. 4 illustrates the impact on the average clustering coefficient and average path length with regard to preference for link formation to a more commonly occurring two-hop neighbor (vertex 8) vis-a-vis preference to a random two-hop neighbor (vertex 6). We notice that the setting up of link 3–8 closes two chains 3–4–8 and 3–7–8 into triangles (contributing to a larger average clustering coefficient of 0.21), but could increase the number of hops from vertices 4, 5 and 6 to the rest of the vertices (contributing to a slightly larger path length of 1.96). On the other hand, setting up of link 3–6 closes only one chain 3–7–6 (contributing to a lower average clustering coefficient of 0.10), but relatively decreases the number of hops on the shortest paths between a majority of the node pairs (an average path length of 1.89). We thus notice that giving preference to the more commonly occurring two-hop neighbor could even double the average clustering coefficient at the expense of a very modest increase (less than 5% increase) in the average path length.

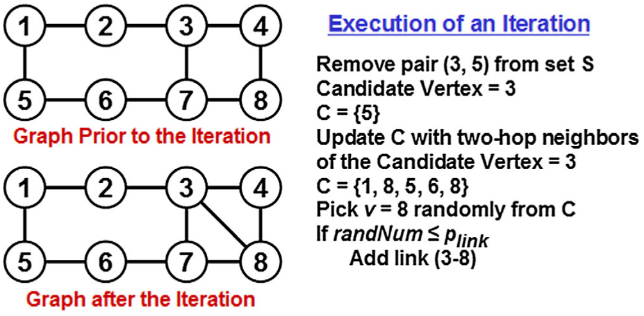
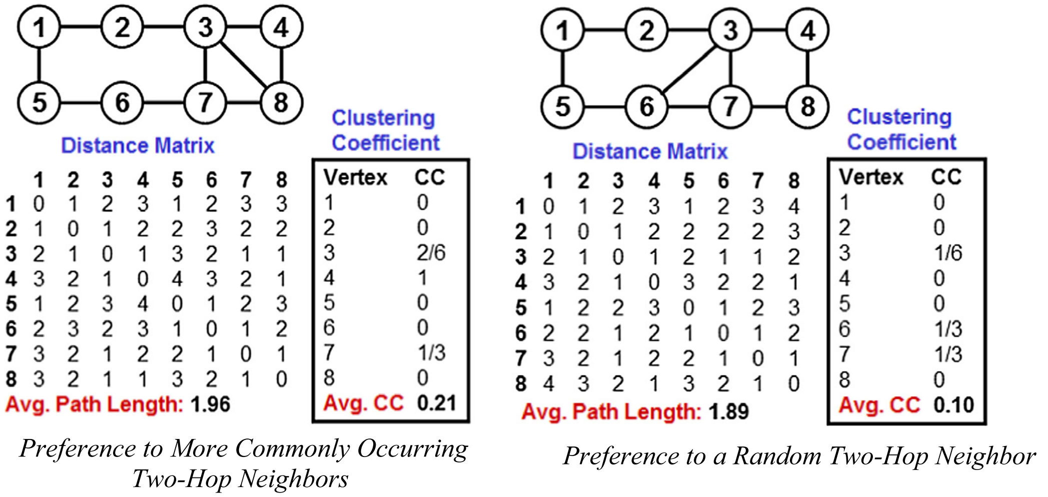

1
2
3
4
5
6
7
8
Graph Prior to the Iteration
1
2
3
4
5
6
7
8
Graph after the Iteration
Execution of an Iteration
Remove pair (3, 5) from set S
Candidate Vertex = 3
C = {5}
Update C with two-hop neighbors
of the Candidate Vertex = 3
C = {1, 8, 5, 6, 8}
Pick v = 8 randomly from C
If randNum ≤ p_link
Add link (3-8)

Figure 3 Example to illustrate the execution of an iteration of the THNP model.

4. Simulations

In this section, we discuss the simulation results observed when the real-world networks are modeled based on the proposed THNP model vis-a-vis the well-known Erdos-Renyi (ER) model for random networks. The common practice in the literature is that if a certain property (like the smaller path length) observed for a real-world network or a theoretical network generated from models (like the THNP model) coincides with that observed for a random network generated from the ER model, then the property is considered to have been observed just by chance in the real-world network or the theoretical network [21,22]. On the other hand, if a certain property (like the inverse relationship between clustering coefficient and node degree) observed for a real-world or theoretical network does not coincide with that observed for an ER-random network, then it confirms the existence of an underlying mechanism (like the two-hop neighbor preference for the THNP model) that could be attributed for the observation of such a property [21,22]. Thus, the ER model is often used in the literature as a theoretical benchmark model for attributing the observation of a certain property to random link formation or preferential link formation. The simulation results presented in this section illustrate that (unlike the ER model) the THNP model exhibits an inverse relationship between clustering coefficient and node degree (attributed to the underlying two-hop neighbor preference mechanism for link formation) as is also observed in several real-world networks.

We evaluate the proximity (similarity) of the theoretical graphs generated with the THNP and ER models with that of the graphs abstracting real-world networks using analytic metrics such as (i) Average Clustering Coefficient-*CC*; (ii) Average Path Length-*PL*; (iii) Average Degree-*K* and (iv) Spectral Radius Ratio for Node Degree-*SRK*. The definitions (including the quantitative formulations) and the procedures to calculate each of the above analytic metrics are given in Section 4.2. The proximity value is modeled as the square root of the sum of the squares of the relative differences with respect to the above analytic metrics for the theoretical graph (abbreviated as *ThG* – could refer to the THNP or ER model graph) and the real-world network graph (abbreviated as *RwG* in Eq. (1)). According to this formulation, the targeted (desirable) proximity value is 0 – indicating that the theoretical graph model captures the fundamental characteristics of a real-world network graph as close as possible. The notations



	1	2	3	4	5	6	7	8
1	0	1	2	3	1	2	3	3
2	1	0	1	2	2	3	2	2
3	2	1	0	1	3	2	1	1
4	3	2	1	0	4	3	2	1
5	1	2	3	4	0	1	2	3
6	2	3	2	3	1	0	1	2
7	3	2	1	2	2	1	0	1
8	3	2	1	1	3	2	1	0

Avg. Path Length: **1.96**

Vertex	CC
1	0
2	0
3	2/6
4	1
5	0
6	0
7	1/3
8	0
Avg. CC	**0.21**

Preference to More Commonly Occurring Two-Hop Neighbors

	1	2	3	4	5	6	7	8
1	0	1	2	3	1	2	3	4
2	1	0	1	2	2	2	2	3
3	2	1	0	1	2	1	1	2
4	3	2	1	0	3	2	2	1
5	1	2	2	3	0	1	2	3
6	2	2	1	2	1	0	1	2
7	3	2	1	2	2	1	0	1
8	4	3	2	1	3	2	1	0

Avg. Path Length: **1.89**

Vertex	CC
1	0
2	0
3	1/6
4	0
5	0
6	1/3
7	1/3
8	0
Avg. CC	**0.10**

Preference to a Random Two-Hop Neighbor

Figure 4 THNP model – impact of the preference to more commonly occurring two-hop neighbor over a random two-hop neighbor.

CC_{ThG}, PL_{ThG}, K_{ThG} and SPR_{ThG} in Eq. (1) indicate respectively the values for the average clustering coefficient, average path length, average degree and spectral radius ratio for node degree incurred with the networks generated from theoretical models such as the THNP model or the ER model. Likewise, the notations CC_{RwG}, PL_{RwG}, K_{RwG} and SPR_{RwG} in Eq. (1) indicate respectively the values for the average clustering coefficient, average path length, average degree and spectral radius ratio for node degree incurred for the real-world network graphs.

$$Proximity\ Value = \sqrt{\left(\frac{CC_{ThG} - CC_{RwG}}{CC_{RwG}}\right)^2 + \left(\frac{PL_{ThG} - PL_{RwG}}{PL_{RwG}}\right)^2 + \left(\frac{K_{ThG} - K_{RwG}}{K_{RwG}}\right)^2 + \left(\frac{SRK_{ThG} - SRK_{RwG}}{SRK_{RwG}}\right)^2} \tag{1}$$

4.1. Erdos-Renyi (ER) model

We simulate the generation of a random graph under the ER model as follows: The input parameters are the number of nodes N and probability of link between any two nodes, p_{link}. We accumulate a set S of all possible pairs of nodes. We proceed in iterations. In each iteration, we randomly remove a pair (u, v) from the set S and generate a random number *randNum* in the range [0…1]. If $randNum \leqslant p_{link}$, we set up the link u-v; otherwise not.

The pseudo code (shown in Fig. 5) for the above procedure of generating a random graph under the ER model could be adapted from the pseudo code described in Fig. 2. The overall time-complexity of the algorithm to generate a random graph of N nodes and L links under the ER model is $\Theta(N^2)$ as it is the time-complexity to run the *for* loop from lines 1–5 plus the time-complexity of the *while* loop from lines 6–13, both of which could be done in $\Theta(N^2)$ time.

4.2. Analytic metrics

In this sub section, we briefly introduce the metrics used in the analysis.

4.2.1. Clustering coefficient

The clustering coefficient of a node is the probability that any two neighbors of a node are connected [3]. Quantitatively, the clustering coefficient of a node is calculated as the ratio of the actual number of links connecting the neighbors of the node divided by the maximum number of possible links between the neighbors of the node. For a node with degree K (i.e., K neighbors), the maximum number of possible links between the neighbors of the node is $K(K-1)/2$. The clustering coefficient of a network (used in formula (1)) is the average of the clustering coefficients of the nodes. The clustering coefficient of a node with an average degree of $K = 2L/N$ could be computed in $\Theta(K^2) = \Theta(L^2/N^2)$ time as it would require to explore the neighborhood of each of the K neighbors of the node. The time-complexity to compute the clustering coefficient of a network would be then $\Theta(NK^2) = \Theta(L^2/N)$.

4.2.2. Path length

The length of a path between two nodes u and v is the minimum number of hops on the shortest path between u and v. The average path length (used in formula (1)) is the average of the path length across all node pairs in the network. The path length of all node pairs could be represented in the form of a distance matrix and could be efficiently computed by running the Floyd's All Pairs Shortest Path algorithm [20] of time-complexity $\Theta(N^3)$ on a graph of N vertices.

4.2.3. Degree

The degree of a node (i.e., the number of neighbors for a node) is the number of links incident on the node. The average degree of a network (used in formula (1)) is the average of the degree values of all the nodes in the network. The degree of a node

```
Input: Number of Nodes, N; Probability of a Link, $p_{link}$
Output: Set of Links, E
Auxiliary Variables: Set of Node Pairs, S
Initialization: $E = \phi$
Begin ER Algorithm
    for every vertex u = 1 to N do
        for every vertex v = u+1 to N do
            $S = S \cup \{(u, v)\}$
        end for
    end for
    while ($S \neq \phi$) do
        Select a pair (u, v) randomly from S
        $S = S - \{(u, v)\}$
        Generate a random number randNum in the range [0...1]
        if $randNum \leq p_{link}$ then
            $E = E \cup \{(u - v)\}$
        end if
    end While
    return E
End ER Algorithm
```

Figure 5 Pseudo code for the Erdos-Renyi (ER) random network generation model.

can be computed in $\Theta(K) = \Theta(L/N)$ time and the average degree of a network of N nodes could be computed in $\Theta(NK) = \Theta(L)$ time.

4.2.4. Spectral radius ratio for node degree

The spectral radius ratio for node degree [23] is the ratio of the principal eigenvalue (a.k.a. spectral radius) of the adjacency matrix [7] of the network graph and the average degree of the nodes in the network. The spectral radius of a graph (denoted $\lambda_{sp}(G)$) is computed using the Power-Iteration method [7] of time-complexity $\Theta(N^3)$ involving at most N iterations; in each iteration, we compute the principal eigenvector of a graph based on its adjacency matrix (A). The principal eigenvector X_{i+1} during the $(i + 1)^{th}$ iteration is given by $AX_{i+1}/||AX_{i+1}||$, where X_i is the principal eigenvector of the graph at the end of the i^{th} iteration and $||AX_i||$ is the normalized value of the product of A and X_i. To start with, $X_i = [1, 1, \ldots, 1]$ – a column vector of all 1 s corresponding to the number of vertices in the graph.

The Power-iteration method (example illustrated in Fig. 6) stops when the normalized value of $||AX_i||$ converges, and the converged normalized value is the spectral radius. If K_{min}, K_{avg} and K_{max} are respectively the minimum, average and maximum values for the node degree, it has been established that $K_{min} \leqslant K_{avg} \leqslant \lambda_{sp}(G) \leqslant K_{max}$ [8]. Accordingly, the spectral radius ratio for node degree (SRK), calculated as $\lambda_{sp}(G)/K_{avg}$, will always be greater than or equal to 1.0.

4.3. Real-world network datasets

In this section, we introduce the six real-world network datasets [9–10] studied in the paper. We consider these six networks to be representatives of real-world networks with a broader range of values for the spectral radius ratio for node degree (i.e., to represent real-world networks with a broader range of values in the variation of node degree). The real-world networks are listed (in the increasing order of their spectral radius ratio for node degree) as follows (the abbreviations used to

Iteration 1

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \\ 2 \end{bmatrix} \equiv \begin{bmatrix} 0.213 \\ 0.426 \\ 0.426 \\ 0.639 \\ 0.426 \end{bmatrix}$$

Normalized Value = 4.69

Iteration 2

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.213 \\ 0.426 \\ 0.426 \\ 0.639 \\ 0.426 \end{bmatrix} = \begin{bmatrix} 0.426 \\ 0.852 \\ 1.065 \\ 1.278 \\ 1.065 \end{bmatrix} \equiv \begin{bmatrix} 0.195 \\ 0.389 \\ 0.486 \\ 0.584 \\ 0.486 \end{bmatrix}$$

Normalized Value = 2.19

Iteration 3

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.195 \\ 0.389 \\ 0.486 \\ 0.584 \\ 0.486 \end{bmatrix} = \begin{bmatrix} 0.389 \\ 0.779 \\ 1.07 \\ 1.361 \\ 1.07 \end{bmatrix} \equiv \begin{bmatrix} 0.176 \\ 0.352 \\ 0.484 \\ 0.616 \\ 0.484 \end{bmatrix}$$

Normalized Value = 2.21

Iteration 4

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.176 \\ 0.352 \\ 0.484 \\ 0.616 \\ 0.484 \end{bmatrix} = \begin{bmatrix} 0.352 \\ 0.792 \\ 1.100 \\ 1.320 \\ 1.100 \end{bmatrix}$$

Normalized Value = 2.21 converges

Figure 6 Example to illustrate the execution of the power iteration method to compute the spectral radius of the adjacency matrix for a network graph.

refer to the networks are indicated in parenthesis, immediately following the network name):

(i) US Football Network (FN): This is a network of 115 teams (representing Division I-A colleges) who played in the Fall 2000 Football season in the US. The nodes represent the teams and there is an edge between two nodes if the corresponding teams played against each other during the season [9].

(ii) US Politics Books Network (PN): This is a network of 105 books related to US Politics sold in Amazon.com. Each book represents a node and there is an edge between two nodes *u* and *v* if someone who bought a book corresponding to node *u* also bought the book corresponding to node *v* and vice versa [9].

(iii) Karate Club Network (KN): This is a social network of 34 members of a karate club at a US university in the 1970s. The members represent the nodes and there is an edge between two nodes if the corresponding members are friends [9].

(iv) C. Elegans Neural Network (EN): This is a network of 297 neurons in the hermaphrodite Caenorhabditis Elegans [9]. Each neuron is a node and there is an edge between two nodes if the corresponding neurons interact with each other (in the form of chemical synapses, gap junctions and neuromuscular junctions).

(v) Les Miserables Network (LN): This is a co-occurrence network of 77 characters in the novel "Les Misèrables" by Victor Hugo. Each character represents a node and there is an edge between two characters if they co-occurred in at least one chapter of the book [9].

(vi) US Airports Network (AN): This is a network of 332 airports in the US during 1997. Each node is an airport and there is an edge between two nodes if there is a direct flight between the two corresponding airports [10].

In Fig. 7, we depict the six real-world networks using the Gephi [11] visualization and analysis tool. The layout algorithm used is the Fruchterman–Reingold layout [12] algorithm. The color of the vertices (varied from white to black: gray scale) is a measure of the clustering coefficient of the vertices – the darker/blacker the color of a vertex, the larger is its clustering coefficient and vice versa. The size of the vertices is a measure of the degree of the vertices – the bigger the size of a vertex, the larger is its degree and vice versa. For all the networks, except the US Football Network (FN), we observe that nodes of larger size (i.e., nodes having a larger degree) are pale (white or close to white) in color; whereas, the nodes of smaller size (i.e., nodes having a smaller degree) are relatively more darker (black) in color. This illustrates the inverse correlation between node degree and clustering coefficient typically observed in real-world networks. A key motivation for the work in this paper is to develop a random network model that can also display such an inverse correlation between node degree and clustering coefficient, unlike the well-known ER model (according to which the clustering coefficient is independent of node degree).

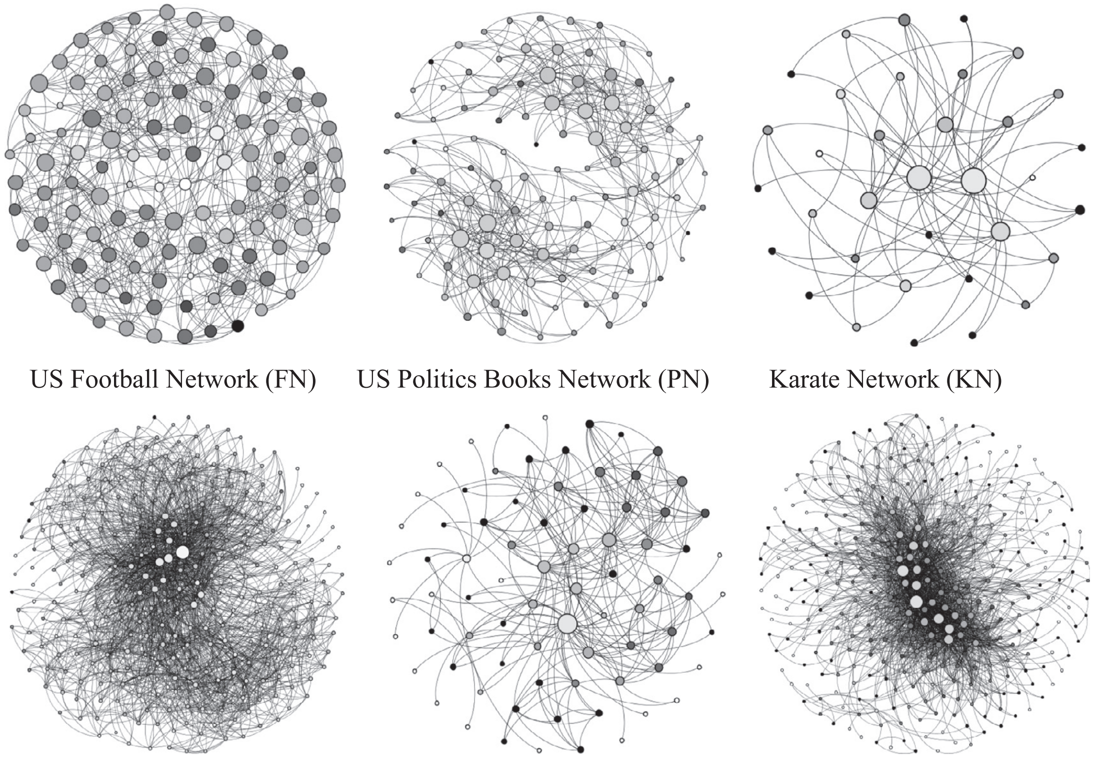
In order to model the real-world network as a random graph under the ER and THNP models, we estimate the probability of link (p_{link}) between any two nodes in the theoretically modeled network to be $\langle K_{RwG}\rangle/N-1$, where $\langle K_{RwG}\rangle$ and N are respectively the average degree of the vertices and the total number of nodes in the real-world network. Fig. 8 presents a comparative look at p_{link} and the analytic metrics with respect to the average degree of the six real-world networks. We observe p_{link} and average correlation coefficient to decrease with increase in the average degree; whereas, the average path length and spectral radius ratio for node degree appear to be almost independent of the average node degree.

4.4. Proximity evaluation

In this section, we evaluate the proximity (similarity) of the theoretically generated random graphs under the THNP and ER models with that of the corresponding real-world networks. We use the estimated probability for a link between any two nodes (estimated as discussed in Section 3.3) and the number of nodes for the real-world networks to generate the corresponding random networks under both the THNP and ER models. For each real-world network (number of nodes and estimated p_{link} values as parameters), we generated 100 instances of the random networks (under each of the two models: THNP and ER) and averaged the values for the analytic metrics. For each real-world network and the corresponding two random networks, we evaluated the analytic metrics (degree, clustering coefficient, path length and spectral radius ratio for node degree) with respect to their average values and distribution, wherever applicable.

Figs. 9 and 10 show respectively the proximity of the average values of the analytic metrics for the random networks under the THNP and ER models in comparison with that obtained under the real-world networks. We observe the THNP model to generate random networks whose average clustering coefficient is very much aligned with that of the corresponding real-world networks (in four of the six cases), whereas, the average clustering coefficient of random networks generated under the ER model is nowhere near the values observed for the real-world networks. We also observe the THNP-model random networks to exhibit relatively larger values for the spectral radius ratio for node degree (compared to the ER model) and incur SPR values that are relatively closer to that obtained for real-world networks. The trade-off is that the average path length of the THNP-model random networks is slightly larger than that obtained for the corresponding real-world networks and the ER-model random networks. The average node degree values for the THNP-model random networks are slightly smaller than the average node degree values for the corresponding real-world networks and the ER-model random networks. This is attributed to the preference for triangle closure using two-hop neighbors (as is observed in the example shown in Fig. 4). Note that triangle closure involving three nodes (say nodes *u*, *v* and *w*) could facilitate a path length of one hop involving any two of these three nodes. We observe that preference for triangle closure (rather than adding the link to arbitrarily connect any two nodes) only increases the chances of observing a larger path length between any two nodes in the network.

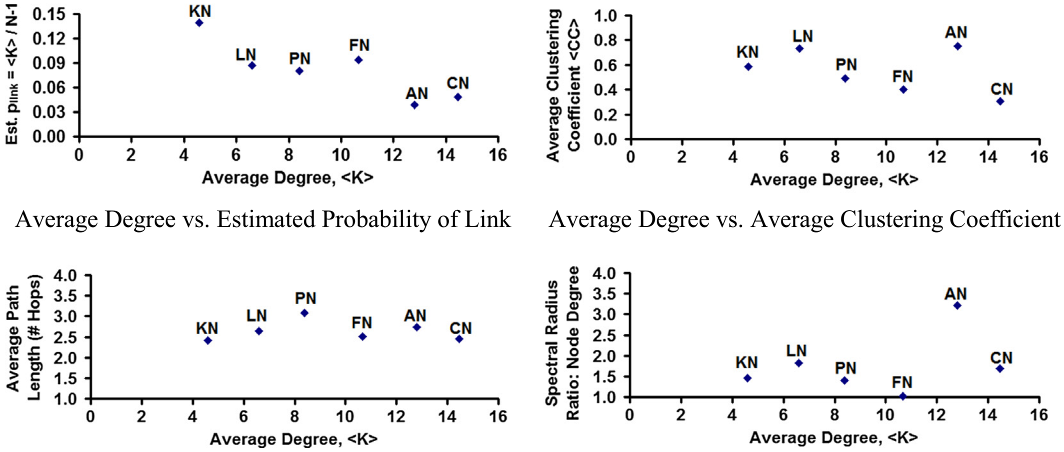
Fig. 11 illustrates a quantitative comparison of the relative proximity of the random networks generated under the THNP and ER models to the corresponding real-world network. The proximity values are computed using Eq. (1) with regard to the average values for the four analytic metrics: degree, clustering coefficient, path length and the spectral radius ratio for node



US Football Network (FN) US Politics Books Network (PN) Karate Network (KN)

C. Elegans Neural Network (CN) Les Misèrables Network (LN) US Airports Network (AN)

Figure 7 Visualization of the real-world networks (degree vs. clustering coefficient).



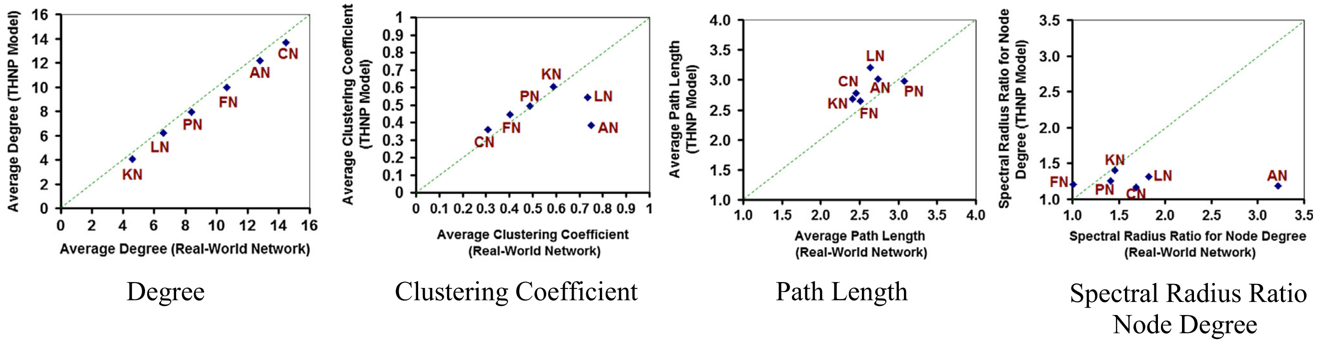
Average Degree vs. Estimated Probability of Link Average Degree vs. Average Clustering Coefficient

Average Degree vs. Average Path Length Average Degree vs. Spectral Radius Ratio: Node Degree

Figure 8 Comparison of the real-world networks based on the estimated probability of link and the analytic metrics.

degree. The ideal proximity value desired is 0; we observe the THNP model to incur a proximity value that is less than 0.5 for five of the six real-world networks analyzed and less than 0.25 for three of the six networks. On the other hand, the ER model incurs a proximity value that is greater than 0.5 for all the six real-world networks analyzed.

The proximity values for the random networks under the ER model are larger than the proximity values for the random networks under the THNP model by factors of 2–7 for five of the six real-world networks. We also observe the THNP model to be of very close proximity to the real-world networks that exhibit a moderate variation in node degree (i.e., for spectral radius ratio for node degree values of the real-world networks in the range of 1.1–1.5). These are the categories of real-world networks (like the Politics Books Network and the Karate Network) that are neither completely random nor completely scale-free, and the ER model or the Barabasi-Albert (BA) model [4] for scale-free networks cannot effectively model them. This is where the proposed THNP model fits the bill. It could effectively model random networks that have a longer



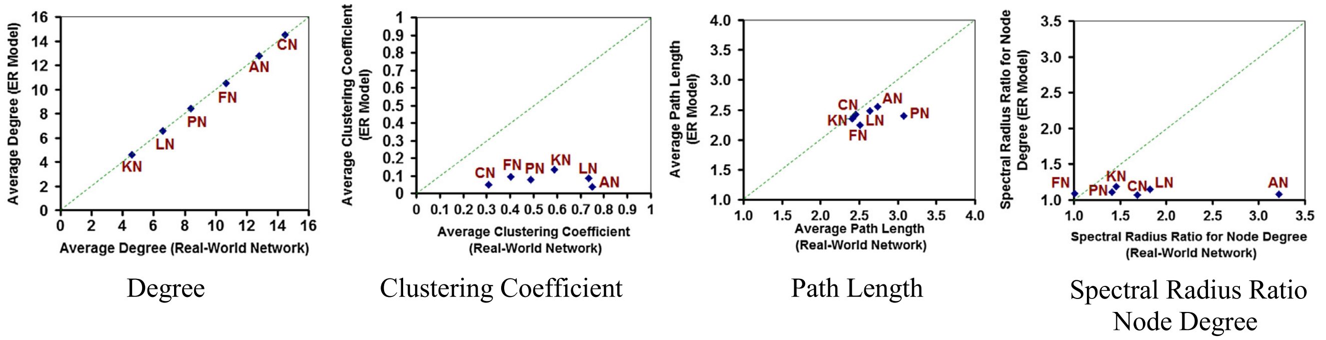
Degree

Clustering Coefficient

Path Length

Spectral Radius Ratio Node Degree

Figure 9 Visualization of the proximity of the real-world networks and the corresponding random networks generated under the THNP model (based on the analytic metrics).



Degree

Clustering Coefficient

Path Length

Spectral Radius Ratio Node Degree

Figure 10 Visualization of the proximity of the real-world networks and the corresponding random networks generated under the ER model (based on the analytic metrics).

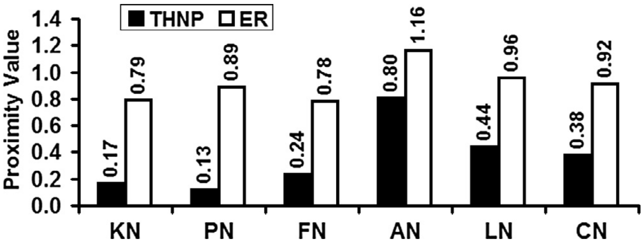


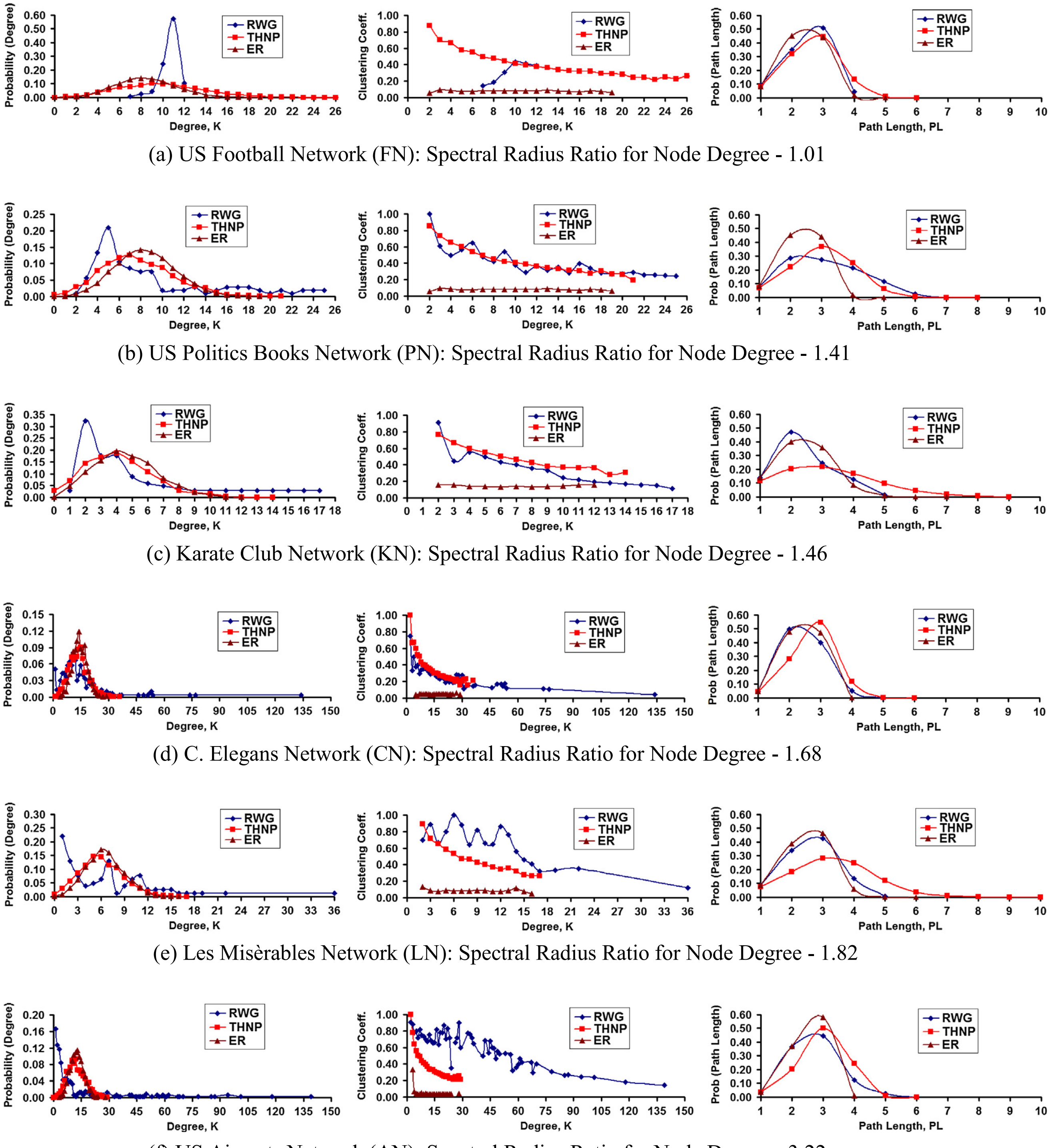
Figure 11 Quantitative evaluation of the proximity of the random networks under the THNP and ER models with that of the corresponding real-world networks (based on the analytic metrics).

tail and/or a longer head (i.e., those networks that look like a combination of random networks and scale-free networks).

Fig. 12(a)–(f) illustrate the distribution of the analytic metrics (degree, clustering coefficient and path length) obtained for the real-world networks (abbreviated as RWG in the figures) and the corresponding random networks generated under the THNP and ER models (computed across all the 100 trials). We obtain the probability distribution for the degree and path length for a particular network graph as follows: We count the number of instances a specific value for the metric is encountered. In the case of degree, the number of instances is the number of nodes exhibiting the particular values of the degree. In the case of path length, the number of instances is the number of node pairs between which the minimum number of hops on a shortest path is the measured path length. The probability of observing a specific value of the analytic metric (degree or path length) for a particular network is then the number of instances that incur the specific value divided by the total number of instances observed.

We observe the degree distribution and path length distribution of the THNP model-based random networks to be relatively more flatter and broader compared to that of the ER model-based random networks. We attribute this to the larger variation in node degree as well as the path length observed in the case of the THNP model-based random networks. Nevertheless, the degree distribution and path length distribution both exhibit a Poisson-style distribution [1], as observed in the ER model-based random networks. Likewise, the THNP-random networks could be construed to still exhibit the small-world property [13], as the average path length observed is only at most 20% more than that observed in the real-world networks and the corresponding ER-model random networks.

To obtain the distribution of degree vs. clustering coefficient, we determine the average of the clustering coefficient of the vertices exhibiting particular values of the node degree. We observe the distribution of the clustering coefficient for the THNP model-based random network graphs to mimic the pattern observed for the real-world networks (i.e., the clustering coefficient decreases with increase in node degree) as well as align closely with the values observed for these networks. On the other hand, as expected, the clustering coefficient of the nodes observed in the ER mode-based random networks is



(a) US Football Network (FN): Spectral Radius Ratio for Node Degree - 1.01

(b) US Politics Books Network (PN): Spectral Radius Ratio for Node Degree - 1.41

(c) Karate Club Network (KN): Spectral Radius Ratio for Node Degree - 1.46

(d) C. Elegans Network (CN): Spectral Radius Ratio for Node Degree - 1.68

(e) Les Misèrables Network (LN): Spectral Radius Ratio for Node Degree - 1.82

(f) US Airports Network (AN): Spectral Radius Ratio for Node Degree - 3.22

Figure 12 Probability distribution of node degree and path length, and node degree vs. clustering coefficient for the real-world networks and random networks under the THNP and ER models.

independent of node degree and simply corresponds to the estimated probability of a link between any two nodes [3].

5. Conclusions

The high-level contributions of this paper are the proposal of a random network graph model that captures the inverse relationship between node degree and clustering coefficient as well as exhibits a relatively larger variation in node degree matching close to real-world networks whose degree distribution exhibits a random network-like Poisson curve with a long tail and/or long head. We accomplish the above by preferring triangle closure for link formation during the evolution of the random network. A node gives preference to attach to its two-hop neighbors (to facilitate triangle closure) rather than to an arbitrary node. Accordingly, we refer to the proposed model as Two-Hop Neighborhood Preference (THNP)-based random network model.

The THNP model is very much suitable for modeling real-world social networks wherein associations (like Friendships in Facebook) are predominantly initiated based on the two-hop neighborhood information (like Friends of Friends as in Facebook). The THNP model also incorporates the feature wherein a node prefers to get connected to a two-hop neighbor with a larger number of common neighbors. This corresponds to a scenario in Facebook wherein a user is more likely to accept a friend request coming from a user with several common friends. Unlike the models discussed in the section on related work, the THNP model does not require any degree sequence of the vertices (including the single-edge degree sequence or triangle degree sequence) that are too time consuming to determine and be used to generate a random graph. The THNP model can be run just with the knowledge of the number of vertices and number of edges in the real-world network graph.

We observe the THNP model-based random networks to exhibit a higher average clustering coefficient and spectral radius ratio for node degree (relatively more closer to that of the real-world networks) and a slightly smaller average node degree. The trade-off is a modest increase in the average path length (at most 20%) compared to those incurred for real-world networks. The proximity values for the THNP model-based random networks to the real-world networks are significantly lower (i.e., the THNP-random networks are relatively more closer to the real-world networks) compared to the proximity values observed for the ER model-based random networks to the real-world networks. To the best of our knowledge, the proposed THNP model is the first such model for random networks wherein the clustering coefficient of the nodes decreases with increase in node degree (as in the case of real-world networks) and still exhibits a Poisson-style degree distribution. On these lines, ours is a significant effort to model real-world networks from a random network point of view.

Acknowledgment

The work was funded through the Massie Chair Grant (#: DE-NA0000654) at Jackson State University, United States.

References

[1] Alon N, Spencer JH. The probabilistic method. 3rd ed. Wiley-Interscience; 2008.

[2] Bollobas B. Random graphs. 2nd ed. Cambridge University Press; 2001.

[3] Newman M. Networks: an introduction. 1st ed. Oxford University Press; 2010.

[4] Barabasi A-L, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509–12.

[5] Erdos P, Renyi A. On random graphs I. Publicationes Mathematicae 1959;6:290–7.

[6] Meghanathan N. Spectral radius as a measure of variation in node degree for complex network graphs. In: Proceedings of the 3rd international conference on digital contents and applications, Hainan, China. p. 30–3.

[7] Lay DC, Lay SR, McDonald JJ. Linear algebra and its applications. 5th ed. Pearson Publishers; 2015.

[8] Butler BK, Siegel PH. Sharp bounds on the spectral radius of nonnegative matrices and digraphs. Linear Algebra Appl 2013;439 ():1468–78.

[9] Newman M. < http://www-personal.umich.edu/~mejn/netdata/ >, [last accessed: November 13, 2015].

[10] Pajek Datasets. < http://vlado.fmf.uni-lj.si/pub/networks/pajek/data/gphs.htm >, [last accessed: November 13, 2015].

[11] Cherven K. Mastering Gephi network visualization. Packt Publishing; 2015.

[12] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software-Practice and Experience 1991;21 ():1129–64.

[13] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393:440–2.

[14] Eggemann N, Noble SD. The clustering coefficient of a scale-free random graph. Discrete Appl Math 2011;159(10):953–65.

[15] Godehardt E, Jaworski J, Rybarczyk K. Random intersection graphs and classification. In: Proceedings of the 30th Annual Conference of Data Analysis and Knowledge Organization, Berlin, Germany. p. 67–74.

[16] Bloznelis M. Degree and clustering coefficient in sparse random intersection graphs. vol. 23, no. 3, pp. 1254–1289, 2013.

[17] Britton T, Deijfen M, Martin-Lof A. Generating simple random graphs with prescribed degree distribution. J Stat Phys 2006;124 ():1377–97.

[18] Newman MEJ. Random graphs with clustering. Phys Rev Lett 2009;103(5):1–5.

[19] Heath LS, Parikh N. Generating random graphs with tunable clustering coefficients. Physica A 2011;390(23–24):4577–87.

[20] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to algorithms. 3rd ed. MIT Press; 2009.

[21] Newman M, Barabasi A-L, Watts DJ. The structure and dynamics of networks. 1st ed. Princeton University Press; 2006.

[22] Barabasi A-L. Network science. 1st ed. Cambridge University Press; 2016.

[23] Meghanathan N. Using spectral radius ratio for node degree to analyze the evolution of complex networks. Int J Comput Netw Commun 2015;7(3):1–12.