Identifying Local Clustering Structures of Evolving Social Networks Using Graph Spectra (Short Paper)

Bo Jiao[1,2], Yiping Bao[1,3](✉), and Jin Wang[1,3]

[1] CETC Big Data Research Institute Co., Ltd., Guiyang 550008, China
[2] School of Mathematics and Big Data, Foshan University, Foshan 528000, China
[3] Guizhou Wingscloud Co., Ltd., Guiyang 550022, Guizhou, China
baoyiping@wingscloud.cn

Abstract. The clustering coefficient has been widely used for identifying the local structure of networks. In this paper, the weighted spectral distribution with 3-cycle (WSD3) that is similar (but not equal) to the clustering coefficient is studied on evolving social networks. It is demonstrated that the ratio of the WSD3 to the network size (i.e., the node number) provides a more sensitive discrimination for the size-independent local structure of social networks in contrast to the clustering coefficient. Moreover, the difference of the WSD3's performances on social networks and communication networks is investigated, and it is found that the difference is induced by the different symmetrical features of the normalized Laplacian spectral densities on these networks.

Keywords: Social networks · Clustering coefficient · Weighted spectral distribution · Normalized Laplacian spectrum

1 Introduction

Many social networks evolve over time, that is, we need some size-independent metrics to identify different structures of these networks. Small-world presents an important structure of social networks, which is commonly indicated by low path length and high clustering coefficient [1]. The clustering coefficient gives information about local connectivity of evolving networks [1], which quantifies the ratio of the number of triangles composed of a given node and its two neighbors to the maximum possible number of these triangles. The weighted spectral distribution (WSD) is defined on the spectrum of the normalized Laplacian matrix and strongly reflects the distribution of random walk N-cycles in a network [2]. If $N = 3$, Fay et al. indicated that the WSD with 3-cycle (WSD3) and the clustering coefficient can be considered to be similar but not equal [2]. However, their work has not quantificationally studied the WSD3 on evolving networks. In this paper, we will investigate the performance of the WSD3 on evolving social networks and analyze the more sensitive discrimination of the WSD3 on social networks with different sizes. Furthermore, we will compare the performance of the WSD3 on social networks to that on communication networks that is useful for our understanding of the difference between these networks in depth.

H. Gao et al. (Eds.): CollaborateCom 2018, LNICST 268, pp. 161–169, 2019.
https://doi.org/10.1007/978-3-030-12981-1_11

Recently, we indicated that the WSD with 4-cycle (WSD4) provides a sensitive discrimination for networks with different average path length [3]. The path length and clustering coefficient are two critical metrics to measure the small-world structure [1]. Hence, the study of this paper will accelerate the WSD's application for evaluating the small-world structure of social networks.

2 Background (Weighted Spectral Distribution)

Social networks can be modeled by a simple and undirected graph *G* = (*V*, *E*) where *V* and *E* respectively denote node set and edge set. Let d_v and *n* respectively denote the degree of node *v* and the number of nodes in *G*. Then, the normalized Laplacian matrix of *G* can be defined as follows [4]:

$$L(G)(u,v) = \begin{cases} 1 & \text{if } u = v \text{ and } d_v \neq 0 \\ -\frac{1}{\sqrt{d_u \cdot d_v}} & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

The normalized Laplacian spectrum is composed of all eigenvalues of $L(G)$: $0 = \lambda_1 \leq \cdots \leq \lambda_n \leq 2$. Please note that these eigenvalues are restricted in the range from 0 to 2 [4]. So, the WSD can be defined as follows [2]:

$$W(G,N) = \sum\nolimits_{i=1,2,\cdots,n} (1-\lambda_i)^N. \quad (2)$$

Let $\Omega \in \{(2(i-1)/k, 2i/k]\}_{i=1}^{k}$ be equally spaced intervals in (0, 2] and $f(\lambda = \theta)_{\theta \in \Omega}$ be the number of eigenvalues falling in interval Ω where $\theta = (2i-1)/k \in \Omega = (2(i-1)/k, 2i/k]$. Then, the WSD can be transformed into [2]:

$$W(G,N) \approx \sum\nolimits_{\lambda=0} (1-\lambda)^N + \sum\nolimits_{\Omega} (1-\theta)^N \cdot f(\lambda=\theta)_{\theta\in\Omega}. \quad (3)$$

When $k \to +\infty$, Eq. (3) goes to Eq. (2).

The multiplicity of the eigenvalue 0 indicates the number of connected components [4], thus only $\lambda_1 = 0$ for the maximum component of a social network. For the maximum component, Eq. (3) can be transformed into:

$$W(G,N) \approx 1 + \sum\nolimits_{\Omega\subseteq(0,1]} (1-\theta)^N \cdot f(\lambda=\theta)_{\theta\in\Omega} + \sum\nolimits_{\Omega\subseteq(1,2]} (1-\theta)^N \cdot f(\lambda=\theta)_{\theta\in\Omega}. \quad (4)$$

Furthermore, Fay et al. demonstrated that the WSD is equal to the sum over all *N*-cycles in *G* [2]:

$$\sum\nolimits_{i=1,2,\cdots,n} (1-\lambda_i)^N = \sum\nolimits_{C} \frac{1}{d_{u_1} d_{u_2} \cdots d_{u_N}}, \quad (5)$$

where $C = u_1 u_2 \ldots u_N$ denotes all *N*-cycles in *G*. Please note that any node in *C* can appear more than once.

3 The WSD3 in a Deterministic Social Network Model

Deterministic models are useful for the rigorous analysis of graph metrics. First, we choose Chen's deterministic social network model [5] to theoretically analyze the WSD3 because the model captures many properties of social networks, such as skipping the levels, small-world, power-law degree distribution and scaling law between clustering coefficient and degree [5].

3.1 Chen's Model

Let T_{n+1} denote the graph of Chen's model, which can be recursively generated with the increasing of hierarchical levels t from 1 to $n+1$: At level $t=1$, the graph is composed of only one node 1_1 (main root). At level $t=2$, two newly added leaf nodes 2_1 and 2_2 are attached to the main root 1_1. At level $t=3$, each of four newly added leaf nodes $3_j(j=1,2,3,4)$ is attached to the main root 1_1 and the subordinate root 2_1 for $j=1,2$ or 2_2 for $j=3,4$. At level $t \in \{4,5,\cdots,n+1\}$, each of 2^{t-1} newly added leaf nodes $t_j(j=1,2,\cdots,2^{t-1})$ is attached to its $t-1$ roots $(t-k)_{\lceil j/2^k \rceil}(k=1,2,\cdots,t-1)$ where $\lceil x \rceil$ rounds the value of x to the nearest integer towards infinity.

3.2 Exact Formula of the WSD3 on Chen's Model

According to Eq. (5), the WSD3 with N = 3 can be calculated as:

$$WSD3 = \sum_{v \in V} WSD3(v), \tag{6}$$

$$WSD3(v) = 2 \cdot \sum_{v_1, v_2 \in N(v) \wedge (v_1, v_2) \in E} \frac{1}{d_v \cdot d_{v_1} \cdot d_{v_2}}, \tag{7}$$

where $N(v)$ denotes the set of all nodes attached to node v in simple and undirected graph $G = (V, E)$. For the 3-circle pattern shown in Fig. 1, there are six corresponding circles, namely ABC, ACB, BAC, BCA, CAB and CBA. Specifically, ABC and ACB are two circles starting from node A. Thus, the constant coefficient 2 in Eq. (6) corresponds to two circles vv_1v_2 and vv_2v_1 that start from node v.

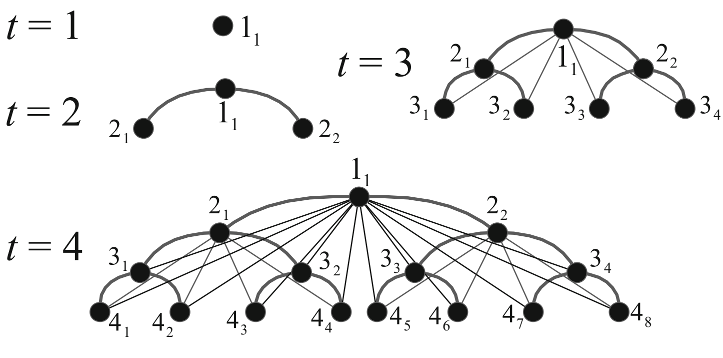

Fig. 1. The generation process of Chen's model.

It is easily to derive the total node number N_{n+1}^T of T_{n+1} and degree d_t of node $t_j (1 \le j \le 2^{t-1})$ at level t in T_{n+1}:

$$N_{n+1}^T = \sum\nolimits_{t=1}^{n+1} 2^{t-1} = 2^{n+1} - 1, \tag{8}$$

$$d_t = 2 \cdot (2^{n-t+1} - 1) + (t-1), \tag{9}$$

At level t, the connection relationship of t_1 is equivalent to other nodes $t_j (2 \le j \le 2^{t-1})$. Thus, we only consider node t_1:

$$N(t_1) = rot(t_1) \cup des(t_1), \tag{10}$$

where $N(t_1)$ includes all nodes attached to node t_1, $rot(t_1) = \{(t-1)_1, (t-2)_1, \cdots, 1_1\}$ is the set of all roots of t_1, and $des(t_1) = \bigcup_{i=1}^{n-t+1} \{(t+i)_1, (t+i)_2, \cdots, (t+i)_{2^i}\}$ is the set of all descendants of t_1. A descendant of t_1 is a node that has a root t_1.

For $\forall k_1 \in rot(t_1)$, we can obtain:

$$N(k_1) \cap N(t_1) = N(t_1)/\{k_1\}, \tag{11}$$

And for $\forall k_j \in des(t_1)$ where k_j is a node at level k,

$$N(k_j) \cap N(t_1) = \left(rot(k_j)/\{t_1\}\right) \cup des(k_j), \tag{12}$$

where

$$rot(k_j) = \left\{(k-1)_{\lceil j/2 \rceil}, (k-2)_{\lceil j/2^2 \rceil}, \cdots, 1_1\right\}, \tag{13}$$

$$des(k_j) = \bigcup_{i=1}^{n-k+1} \left\{(k+i)_{2^i \cdot (j-1)+1}, (k+i)_{2^i \cdot (j-1)+2}, \cdots, (k+i)_{2^i \cdot j}\right\}, \tag{14}$$

According to Eq. (7) and Eqs. (10)–(14), we can obtain:

$$WSD3(t_1) = \frac{1}{d_t}\left[\sum_{k=1}^{t-1} \frac{1}{d_k}\left(\sum_{i=1}^{t-1} \frac{1}{d_i} - \frac{1}{d_k} + \sum_{i=1}^{n-t+1} \frac{2^i}{d_{t+i}}\right) + \sum_{k=t+1}^{n+1} \frac{2^{k-t}}{d_k}\left(\sum_{i=1}^{k-1} \frac{1}{d_i} - \frac{1}{d_t} + \sum_{i=1}^{n-k+1} \frac{2^i}{d_{k+i}}\right)\right], \tag{15}$$

There are 2^{t-1} nodes at level t. Thus, based on Eq. (6),

$$WSD3 = \sum\nolimits_{t=1}^{n+1} 2^{t-1} \cdot WSD3(t_1), \tag{16}$$

Theorem 1. When $n \to +\infty$, for $\forall \gamma > 1 \wedge 1 \leq t \leq n+1$,

$$2^{n-t+1} \leq d_t \leq 2^{\gamma \cdot n-t+3}, \tag{17}$$

where d_t is defined by Eq. (9).

Proof. As is well known, $d_t \geq 2^{n-t+1} + (2^{n-t+1} - 2)$ if $t \geq 1$. When $1 \leq t \leq n$, $d_t \geq 2^{n-t+1}$. When $t = n+1$, $d_t = n \wedge 2^{n-t+1} = 1 \Rightarrow d_t \geq 2^{n-t+1}$. When $n \to +\infty$, for $\forall \gamma > 1$, $2^{(\gamma-1)\cdot n} \geq n$. Thus, when $n \to +\infty$, for $\forall \gamma > 1 \wedge 1 \leq t \leq n+1$, $t-1 \leq n \leq 2^{(\gamma-1)\cdot n} \leq 2^{\gamma \cdot n-t+1}$. Moreover, we can determine $d_t \leq 3 \cdot 2^{\gamma \cdot n-t+1} \leq 2^{\gamma \cdot n-t+3}$. The proof is completed. □

Additionally, when $n \to +\infty$, with $\gamma \to 1$, we can obtain:

$$2^{n-t+1} \leq d_t \leq 2^{n-t+3}, \tag{18}$$

Using Eqs. (8), (15), (16) and (18), we can determine:

$$\frac{3}{80} \leq \lim_{n \to +\infty} \frac{WSD3}{N_{n+1}^T} \leq \frac{12}{5}, \tag{19}$$

Therefore, the WSD3 of T_{n+1} steadily increases as network size N_{n+1}^T grows.

4 The WSD3 in a Stochastic Social Network Model

Deterministic models cannot capture randomized procedure, so we choose Leskovec's stochastic social network model [6] to numerically analyze the WSD3.

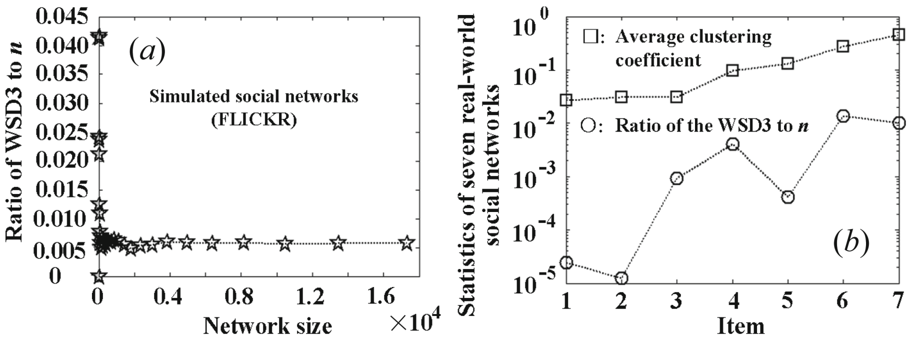

Fig. 2. Results on simulated and real-world social networks. (*a*) WSD3/*n* vs. *n* of Leskovec's model (where *n* is the network size). (*b*) Comparisons of average clustering coefficient and WSD3/*n* on seven real-world social networks (the items of *x*-axis are shown in Table 1).

Leskovec's model [6] has four inputs (namely, $N(t)$, λ, α and β) and three critical processes (namely, node arrival, edge initiation and edge destination selection). The model constructs networks using a recursive method. Specifically, $N(t)$ denotes the number of nodes arriving at step t, λ denotes an exponential distribution parameter predefined for sampling the lifetime of newly arrived nodes, and α, β are two parameters predefined for a special distribution which is used for sampling time gaps between two edge initiation processes of a node. Also, the edge destination selection process is defined by a random-random triangle-closing method. If $N(t) = 0.25^t$, $\lambda = 0.0092$, $\alpha = 0.84$ and $\beta = 0.002$, Leskovec et al. [6] numerically confirmed that the model can generate an evolving system similar to the real-world social network FLICKR (flickr.com). Simulated social networks FLICKR constructed at iteration steps from 1 to 33 are considered in Fig. 2(a). Note that, at step 33, the maximum network size is $\sum_{t=1,2,\ldots,33} <N(t)> \; = 17,300$ where $<N(t)>$ rounds towards the nearest integer of $N(t)$. As shown in Fig. 2(a), after 11 iteration steps, when $n > 65$, the ratio of the WSD3 to n tends towards a positive constant that is consistent with Eq. (19). In other words, the ratio WSD3/n is a size-independent metric which can be used for the comparison of social networks with different sizes.

5 The WSD3 in Real-World Social Networks

To analyze the applicability of the WSD3 in realistic social networks, we choose seven real-world networks with different sizes from the Stanford Large Network Dataset Collection [7], as shown in Table 1. Their network sizes span the range from 4,039 to 2,394,385. We sort these networks by the order of increasing average clustering coefficient. The performances of the average clustering coefficient and the WSD3/n in the real-world social networks listed in Table 1 are shown in Fig. 2(b), where the average clustering coefficient monotonically grows with increasing item.

Table 1. Real-world social networks.

Item	Description	Node number	Edge number
1	Wikipedia Talk network	2,394,385	5,021,410
2	EU email network	265,214	420,045
3	Slashdot social network, Nov. 2008	77,360	905,468
4	Brightkite social network	58,228	214,078
5	Wikipedia vote network	7,115	103,689
6	Enron email network	36,692	183,831
7	Social circles: Facebook	4,039	88,234

According to the comparisons of Fig. 2(b), we can find that the WSD3/n does not monotonically grows and the WSD3/n can commonly provide more sensitive discrimination for the seven different social networks. Both the WSD3/n and the average clustering coefficient pay attention on the triangle structure, but the former is different from the latter in many situations. Based on Eq. (7), the WSD3 located in a given node v quantifies the probability of leaving and returning the node v through two middle nodes for a random walker.

At the same time, we can give the formula of the clustering coefficient located a given node v as follows:

$$ACC(v) = \frac{En(v)}{Total(v)} = 2 \cdot \sum_{v_1, v_2 \in N(v) \wedge (v_1, v_2) \in E} \frac{1}{d_v \cdot (d_v - 1)}, \tag{20}$$

where

$$En(v) = \sum_{v_1, v_2 \in N(v) \wedge (v_1, v_2) \in E} 1, \quad Total(v) = \frac{d_v(d_v - 1)}{2} \tag{21}$$

As is well known, $En(v)$ is the number of links between two nodes attached to the node v, and $Total(v)$ denotes the maximum possible number of links between nodes in $N(v)$. The information of node degrees d_{v1} and d_{v2} are included in Eq. (7) but not included in Eq. (20). Also, Eq. (20) shows that the clustering coefficient located in a given node v only reflects the probability of $En(v)$ links existing between two nodes attached to the node v. Hence, the WSD3/n can provide more sensitive discrimination for social networks in contrast to the average clustering coefficient in general that is consistent with the phenomena shown in Fig. 2(*b*).

6 Difference of the WSD3 on Social and Communication Networks

Existing works [3] indicated that the normalized Laplacian spectral density is quasi-symmetric about one on many communication networks, such as the Interdomain Internet topology. Based on Eq. (4), the quasi-symmetry of the spectral density provides the main reason of that the WSD3/n tends towards zero when n goes to infinity (i.e., the WSD3 almost does not grow with increasing n).

Positive-Feedback Preference (PFP) [8] is an evolving model for Interdomain Internet topologies. The model has three inputs (namely, p, q and δ). Specifically, p and q are two probabilities used to select three different growth mechanisms at the evolving process, and δ is predefined for determining the preferential attachment rule. Zhou et al. [8] numerically found that the model is similar to the realistic Internet system when $p = 0.3$, $q = 0.1$ and $\delta = 0.048$.

We use the PFP model to simulate Internet topologies with increasing network size from 2,000 to 12,000. The WSD3 vs. n of these topologies is exhibited in Fig. 3(*a*), which shows that the WSD3 of the Internet topology does not grow as n increases. This result is obviously different from that exhibited in Fig. 2(*a*) (namely, the WSD3 grows sublinearly with increasing n).

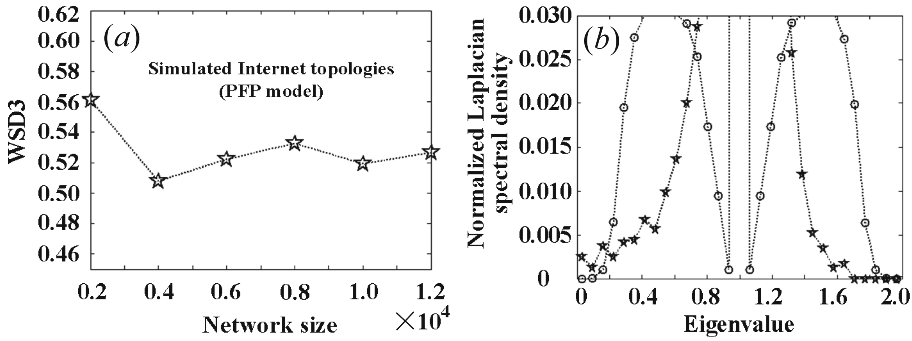

Fig. 3. Difference of the WSD3 on social and communication networks. (*a*) WSD3 vs. *n* of PFP model (where *n* is the network size). (*b*) Comparisons of spectral densities on social and communication networks. ☆: the social network of Item 7 listed in Table 1. O: the Internet topology simulated by the PFP model (having 12,000 nodes).

All eigenvalues of the normalized Laplacian spectrum fall in the range from 0 to 2 [4], so we decompose the range into 31 equally spaced intervals and evaluate the distribution of the eigenvalues falling in the 31 intervals, as shown in Fig. 3(*b*). From Fig. 3(*b*), we can confirm that the spectral density of the Internet topology is quasi-symmetric about one, whereas the spectral density of the social network is obviously asymmetric. Based on Eq. (4), the WSD3 (with $N = 3$) is not far away from one if the spectral density is quasi-symmetric about one because the sum of second and third terms of the right-hand side of Eq. (4) is close to zero. Thus, the asymmetric spectral density of social networks is the main reason for that the WSD3/*n* tends towards a positive constant as *n* goes to infinity.

7 Conclusion

Triangles are important for identifying the local clustering structure of networks. In this paper, we study the performance of a graph spectral metric (i.e., the WSD3) which is defined on the 3-cycle structure. In contrast to the average clustering coefficient, we indicate that the ratio of the WSD3 to network size can provide more sensitive discrimination for the size-independent structure of evolving social networks. Moreover, we compare the performances of the WSD3 on social and communication networks and find that the asymmetric spectral density can provide an effective interpretation for that the WSD3 grows sublinearly with increasing network size. The contributions of this paper are useful for the comparison of social networks with different sizes and are important for our deep understanding of the local clustering structure of evolving social networks.

Acknowledgement. This research has been supported by the Open Fund Project of National Engineering Laboratory for Big Data Application on Improving Government Governance Capabilities.

References

1. Paluck, E.L., Shepherd, H., Aronow, P.M.: Changing climates of conflict: a social network experiment in 56 schools. Proc. Natl. Acad. Sci. **113**(3), 566–571 (2016)
2. Fay, D., Haddadi, H., Thomason, A., et al.: Weighted spectral distribution for internet topology analysis: theory and applications. IEEE/ACM Trans. Netw. **18**(1), 164–176 (2010)
3. Jiao, B., Shi, J., Wu, X., et al.: Correlation between weighted spectral distribution and average path length in evolving networks. Chaos Interdisc. J. Nonlinear Sci. **26**(2), 023110 (2016)
4. Xie, P., Zhang, Z., Comellas, F.: The normalized Laplacian spectrum of subdivisions of a graph. Appl. Math. Comput. **286**, 250–256 (2016)
5. Chen, M., Yu, B., Xu, P., et al.: A new deterministic complex network model with hierarchical structure. Phys. A Stat. Mech. Appl. **385**(2), 707–717 (2007)
6. Leskovec, J., Backstrom, L., Kumar, R., et al.: Microscopic evolution of social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 462–470. ACM (2008)
7. Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/. Accessed 19 July 2018
8. Zhou, S., Mondragón, R.J.: Accurately modeling the Internet topology. Phys. Rev. E **70**(6), 066108 (2004)