Local Clustering Coefficient in Generalized Preferential Attachment Models

Alexander Krot[1(✉)] and Liudmila Ostroumova Prokhorenkova[2,3]

[1] Moscow Institute of Physics and Technology, Moscow, Russia
al.krot.kav@gmail.com
[2] Yandex, Moscow, Russia
[3] Moscow State University, Moscow, Russia

Abstract. In this paper, we analyze the local clustering coefficient of preferential attachment models. A general approach to preferential attachment was introduced in [19], where a wide class of models (PA-class) was defined in terms of constraints that are sufficient for the study of the degree distribution and the clustering coefficient. It was previously shown that the degree distribution in all models of the PA-class follows a power law. Also, the global clustering coefficient was analyzed and a lower bound for the average local clustering coefficient was obtained. We expand the results of [19] by analyzing the local clustering coefficient for the PA-class of models. Namely, we analyze the behavior of $C(d)$ which is the average local clustering for the vertices of degree d.

Keywords: Networks · Random graph models · Preferential attachment · Clustering coefficient

1 Introduction

Nowadays there are a lot of practical problems connected with the analysis of growing real-world networks, from Internet and society networks [1,6,9] to biological networks [2]. Models of real-world networks are used in physics, information retrieval, data mining, bioinformatics, etc. An extensive review of real-world networks and their applications can be found elsewhere (e.g., see [1,6,7,13]).

It turns out that many real-world networks of diverse nature have some typical properties: small diameter, power-law degree distribution, high clustering, and others [15,17,18,24]. Probably the most extensively studied property of networks is their vertex degree distribution. For the majority of studied real-world networks, the portion of vertices with degree d was observed to decrease as $d^{-\gamma}$, usually with $2 < \gamma < 3$ [3–6,10,14].

Another important characteristic of a network is its clustering coefficient, which has the following two most used versions: the global clustering coefficient and the average local clustering coefficient (see Sect. 2.3 for the definitions). It is believed that for many real-world networks both the average local and the global clustering coefficients tend to non-zero limit as the network becomes large.

D.F. Gleich et al. (Eds.): WAW 2015, LNCS 9479, pp. 15–28, 2015.
DOI: 10.1007/978-3-319-26784-5_2

Indeed, in many observed networks the values of both clustering coefficients are considerably high [18].

The most well-known approach to modeling complex networks is the preferential-attachment idea. Many different models are based on this idea: LCD [8], Buckley-Osthus [11], Holme-Kim [16], RAN [25], and many others. A general approach to preferential attachment was introduced in [19], where a wide class of models was defined in terms of constraints that are sufficient for the study of the degree distribution (PA-class) and the clustering coefficient (T-subclass of PA-class).

In this paper, we analyze the behavior of $C(d)$ — the average local clustering coefficient for the vertices of degree d — in the T-subclass. It was previously shown that in real-world networks $C(d)$ usually decreases as $d^{-\psi}$ with some parameter $\psi > 0$ [12,21,23]. For some networks, $C(d)$ scales as a power law $C(d) \sim d^{-1}$ [13,20]. In the current paper, we prove that in *all* models of the T-subclass the local clustering coefficient $C(d)$ asymptotically behaves as $C \cdot d^{-1}$, where C is some constant.

The remainder of the paper is organized as follows. In Sect. 2, we give a formal definition of the PA-class and present some known results. Then, in Sect. 3, we state new results on the behavior of local clustering $C(d)$. We prove the theorems in Sect. 4. Section 5 concludes the paper.

2 Generalized Preferential Attachment

2.1 Definition of the PA-class

In this section, we define the PA-class of models which was first suggested in [19]. Let G_m^n $(n \geq n_0)$ be a graph with n vertices $\{1, \ldots, n\}$ and mn edges obtained as a result of the following process. We start at the time n_0 from an arbitrary graph $G_m^{n_0}$ with n_0 vertices and mn_0 edges. On the $(n+1)$-th step $(n \geq n_0)$, we make the graph G_m^{n+1} from G_m^n by adding a new vertex $n+1$ and m edges connecting this vertex to some m vertices from the set $\{1, \ldots, n, n+1\}$. Denote by d_v^n the degree of a vertex v in G_m^n. If for some constants A and B the following conditions are satisfied

$$\mathsf{P}\left(d_v^{n+1} = d_v^n \mid G_m^n\right) = 1 - A\frac{d_v^n}{n} - B\frac{1}{n} + O\left(\frac{(d_v^n)^2}{n^2}\right), \ 1 \leq v \leq n, \tag{1}$$

$$\mathsf{P}\left(d_v^{n+1} = d_v^n + 1 \mid G_m^n\right) = A\frac{d_v^n}{n} + B\frac{1}{n} + O\left(\frac{(d_v^n)^2}{n^2}\right), \ 1 \leq v \leq n, \tag{2}$$

$$\mathsf{P}\left(d_v^{n+1} = d_v^n + j \mid G_m^n\right) = O\left(\frac{(d_v^n)^2}{n^2}\right), \ 2 \leq j \leq m, \ 1 \leq v \leq n, \tag{3}$$

$$\mathsf{P}(d_{n+1}^{n+1} = m + j) = O\left(\frac{1}{n}\right), \ 1 \leq j \leq m\,, \tag{4}$$

then the random graph process G_m^n is a model from the PA-class. Here, as in [19], we require $2mA + B = m$ and $0 \le A \le 1$.

As it is explained in [19], even fixing values of parameters A and m does not specify a concrete procedure for constructing a network. There are a lot of models possessing very different properties and satisfying the conditions (1–4), e.g., the LCD, the Buckley–Osthus, the Holme–Kim, and the RAN models.

2.2 Power Law Degree Distribution

Let $N_n(d)$ be the number of vertices of degree d in G_m^n. The following theorems on the expectation of $N_n(d)$ and its concentration were proved in [19].

Theorem 1. *For every model in PA-class and for every $d \ge m$*

$$\mathsf{E}N_n(d) = c(m,d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right),$$

where

$$c(m,d) = \frac{\Gamma\left(d+\frac{B}{A}\right)\Gamma\left(m+\frac{B+1}{A}\right)}{A\,\Gamma\left(d+\frac{B+A+1}{A}\right)\Gamma\left(m+\frac{B}{A}\right)} \overset{d\to\infty}{\sim} \frac{\Gamma\left(m+\frac{B+1}{A}\right) d^{-1-\frac{1}{A}}}{A\,\Gamma\left(m+\frac{B}{A}\right)}$$

and $\Gamma(x)$ is the gamma function.

Theorem 2. *For every model from the PA-class and for every $d = d(n)$ we have*

$$\mathsf{P}\left(|N_n(d) - \mathsf{E}N_n(d)| \ge d\sqrt{n}\,\log n\right) = O\left(n^{-\log n}\right).$$

Therefore, for any $\delta > 0$ there exists a function $\varphi(n) \in o(1)$ such that

$$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \le n^{\frac{A-\delta}{4A+2}} : |N_n(d) - \mathsf{E}N_n(d)| \ge \varphi(n)\,\mathsf{E}N_n(d)\right) = 0.$$

These two theorems mean that the degree distribution follows (asymptotically) the power law with the parameter $1 + \frac{1}{A}$.

2.3 Clustering Coefficient

A T-subclass of the PA-class was introduced in [19]. In this case, the following additional condition is required:

$$\mathsf{P}\left(d_i^{n+1} = d_i^n + 1, d_j^{n+1} = d_j^n + 1 \mid G_m^n\right) = e_{ij}\frac{D}{mn} + O\left(\frac{d_i^n d_j^n}{n^2}\right). \tag{5}$$

Here e_{ij} is the number of edges between vertices i and j in G_m^n and D is a positive constant. Note that this property still does not define the correlation between edges completely, but it is sufficient for studying both global and average local clustering coefficients.

Let us now define the clustering coefficients. The *global clustering coefficient* $C_1(G)$ is the ratio of three times the number of triangles to the number of pairs of

adjacent edges in G. The *average local clustering coefficient* is defined as follows: $C_2(G) = \frac{1}{n}\sum_{i=1}^{n} C(i)$, where $C(i)$ is the local clustering coefficient for a vertex i: $C(i) = \frac{T^i}{P_2^i}$, where T^i is the number of edges between neighbors of the vertex i and P_2^i is the number of pairs of neighbors. Note that both clustering coefficients are defined for graphs without multiple edges.

The following theorem on the global clustering coefficient in the T-subclass was proven in [19].

Theorem 3. *Let G_m^n belong to the T-subclass with $D > 0$. Then, for any $\varepsilon > 0$*

(1) If $2A < 1$, then ***whp*** $\frac{6(1-2A)D-\varepsilon}{m(4(A+B)+m-1)} \le C_1(G_m^n) \le \frac{6(1-2A)D+\varepsilon}{m(4(A+B)+m-1)}$;
(2) If $2A = 1$, then ***whp*** $\frac{6D-\varepsilon}{m(4(A+B)+m-1)\log n} \le C_1(G_m^n) \le \frac{6D+\varepsilon}{m(4(A+B)+m-1)\log n}$;
(3) If $2A > 1$, then ***whp*** $n^{1-2A-\varepsilon} \le C_1(G_m^n) \le n^{1-2A+\varepsilon}$.

Theorem 3 shows that in some cases ($2A \ge 1$) the global clustering coefficient $C_1(G_m^n)$ tends to zero as the number of vertices grows.

The average local clustering coefficient $C_2(G_m^n)$ was not fully analyzed previously, but it was shown in [19] that $C_2(G_m^n)$ does not tend to zero for the T-subclass with $D > 0$. In the next section, we fully analyze the behavior of the average local clustering coefficient for the vertices of degree d.

3 The Average Local Clustering for the Vertices of Degree *d*

In this section, we analyze the asymptotic behavior of $C(d)$ — the average local clustering for the vertices of degree d. Let $T_n(d)$ be the number of triangles on the vertices of degree d in G_m^n (i.e., the number of edges between the neighbors of the vertices of degree d). Then, $C(d)$ is defined in the following way:

$$C(d) = \frac{T_n(d)}{N_n(d)\binom{d}{2}}. \tag{6}$$

In other words, $C(d)$ is the local clustering coefficient averaged over all vertices of degree d. In order to estimate $C(d)$ we should first estimate $T_n(d)$. After that, we can use Theorems 1 and 2 on the behavior of $N_n(d)$.

We prove the following result on the expectation of $T_n(d)$.

Theorem 4. *Let G_m^n belong to the T-subclass of the PA-class with $D > 0$. Then*

(1) if $2A < 1$, then $\mathsf{E}T_n(d) = K(d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right)$;
(2) if $2A = 1$, then $\mathsf{E}T_n(d) = K(d)\left(n + O\left(d^{2+\frac{1}{A}} \cdot \log(n)\right)\right)$;
(3) if $2A > 1$, then $\mathsf{E}T_n(d) = K(d)\left(n + O\left(d^{2+\frac{1}{A}} \cdot n^{2A-1}\right)\right)$;

where $K(d) = c(m,d)\left(D + \frac{D}{m}\cdot\sum_{i=m}^{d-1}\frac{i}{Ai+B}\right) \overset{d\to\infty}{\sim} \frac{D}{A\,m}\cdot\frac{\Gamma\left(m+\frac{B+1}{A}\right)}{A\,\Gamma\left(m+\frac{B}{A}\right)}\cdot d^{-\frac{1}{A}}$.

Second, we show that the number of triangles on the vertices of degree d is highly concentrated around its expectation.

Theorem 5. *Let G_m^n belong to the T-subclass of the PA-class with $D > 0$. Then for every $d = d(n)$*

(1) if $2A < 1$: $\mathsf{P}\left(|T_n(d) - \mathsf{E}T_n(d)| \geq d^2 \sqrt{n} \log n\right) = O\left(n^{-\log n}\right)$;
(2) if $2A = 1$: $\mathsf{P}\left(|T_n(d) - \mathsf{E}T_n(d)| \geq d^2 \sqrt{n} \log^2 n\right) = O\left(n^{-\log n}\right)$;
(3) if $2A > 1$: $\mathsf{P}\left(|T_n(d) - \mathsf{E}T_n(d)| \geq d^2 n^{2A-\frac{1}{2}} \log n\right) = O\left(n^{-\log n}\right)$.

Consequently, for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that

(1) if $2A \leq 1$: $\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A-\delta}{4A+2}} : |T_n(d) - \mathsf{E}T_n(d)| \geq \varphi(n)\, \mathsf{E}T_n(d)\right) = 0$;
(2) if $2A > 1$:
$\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A(3-4A)-\delta}{4A+2}} : |T_n(d) - \mathsf{E}T_n(d)| \geq \varphi(n)\, \mathsf{E}T_n(d)\right) = 0$.

As a consequence of Theorems 1, 2, 4, and 5, we get the following result on the average local clustering coefficient $C(d)$ for the vertices of degree d in G_m^n.

Theorem 6. *Let G_m^n belong to the T-subclass of the PA-class. Then for any $\delta > 0$ there exists a function $\varphi(n) = o(1)$ such that*

(1) if $2A \leq 1$: $\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A-\delta}{4A+2}} : \left|C(d) - \frac{K(d)}{\binom{d}{2} c(m,d)}\right| \geq \frac{\varphi(n)}{d}\right) = 0$;
(2) if $2A > 1$: $\lim_{n\to\infty} \mathsf{P}\left(\exists\, d \leq n^{\frac{A(3-4A)-\delta}{4A+2}} : \left|C(d) - \frac{K(d)}{\binom{d}{2} c(m,d)}\right| \geq \frac{\varphi(n)}{d}\right) = 0$.

Note that $\frac{K(d)}{\binom{d}{2} c(m,d)} = \frac{2D}{d(d-1)m}\left(m + \sum_{i=m}^{d-1} \frac{i}{Ai+B}\right) \overset{d\to\infty}{\sim} \frac{2D}{mA} \cdot d^{-1}$.

It is important to note that Theorems 5 and 6 are informative only for $A < \frac{3}{4}$, since only in this case the value $n^{\frac{A(3-4A)-\delta}{4A+2}}$ grows.

In the next section, we first prove Theorem 4. Then, using the Azuma–Hoeffding inequality, we prove Theorem 5. Theorem 6 is a corollary of Theorems 1, 2, 4, and 5.

4 Proofs

In all the proofs we use the notation $\theta(\cdot)$ for error terms. By $\theta(X)$ we denote an arbitrary function such that $|\theta(X)| < X$.

4.1 Proof of Theorem 4

We need the following auxiliary theorem.

Theorem 7. *Let W_n be the sum of the squares of the degrees of all vertices in a model from the PA-class. Then*

(1) if $2A < 1$, *then* $\mathsf{E}W_n = O(n)$,
(2) if $2A = 1$, *then* $\mathsf{E}W_n = O(n \cdot \log(n))$,
(3) if $2A > 1$, *then* $\mathsf{E}W_n = O(n^{2A})$.

This statement is mentioned in [19] and it can be proved by induction. Also, let $S(n, d)$ be the sum of the degrees of all the neighbors of all vertices of degree d. Note that $S(n, d)$ is not greater than the sum of the degrees of the neighbors of all vertices. The last is equal to W_n, because each vertex of degree d adds d^2 to the sum of the degrees of the neighbors of all vertices. So, for any d we have

$$\mathsf{E}S(n, d) \le \mathsf{E}W_n. \tag{7}$$

Now we can prove Theorem 4. Note that we do not take into account the multiplicities of edges when we calculate the number of triangles, since the clustering coefficient is defined for graphs without multiple edges. This does not affect the final result since the number of multiple edges is small for graphs constructed according to the model [7].

We prove the statement of Theorem 4 by induction on d. Also, for each d we use induction on n. First, consider the case $d = m$. The expected number of triangles on any vertex t of degree m is equal to $\mathsf{E}\sum_{(i,j)\in E(G_m^t)} \left(e_{ij}\frac{D}{mt} + O\left(\frac{d_i^t d_j^t}{t^2}\right)\right)$ (see (5)). As G_m^t has exactly mt edges, we get $\mathsf{E}\sum_{(i,j)\in E(G_m^t)} \left(e_{ij}\frac{D}{mt} + O\left(\frac{d_i^t d_j^t}{t^2}\right)\right) = D + o(1)$. The fact that $\mathsf{E}\sum_{(i,j)\in E(G_m^t)} O\left(\frac{d_i d_j}{t^2}\right) = O\left(\frac{\mathsf{E}W_t}{t^2}\right) = o(1)$ can be shown by induction using the conditions (1–4). We also know (see Theorem 1) that $\mathsf{E}N_n(m) = c(m, m)\, n + O(1)$. So, $\mathsf{E}T_n(m) = (D + o(1))\,(c(m, m)\, n + O(1)) = K(m)\,(n + O(1))$. This concludes the proof for the case $d = m$ for all values of A ($2A < 1$, $2A = 1$ and $2A > 1$).

Consider the case $d > m$. Note that the number of triangles on a vertex of degree d is $O(d)$, since this number is $O(1)$ when this vertex appears plus at each step we get a triangle only if we hit both the vertex under consideration and a neighbor of this vertex, and our vertex degree equals d, therefore we get at most $d\,m$ triangles. Also, $\mathsf{E}N_n(d) = c(m, d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right)$. So we have $\mathsf{E}T_n(d) = O(d)\, c(m, d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right)$. In particular, for $n \le Q \cdot d^2$ (where the constant Q depends only on A and m and will be defined later) we have $\mathsf{E}T_n(d) = O\left(c(m, d)\, d^{3+\frac{1}{A}}\right) = O\left(d^2\right) = K(d) \cdot O\left(d^{2+\frac{1}{A}}\right)$. This concludes the proof for the case $d > m$, $n \le Qd^2$ for all values of A.

Now, consider the case $d > m$, $n > Q\,d^2$. Once we add a vertex $n + 1$ and m edges, we have the following possibilities.

1. At least one edge hits a vertex of degree d. Then $T_n(d)$ is decreased by the number of triangles on this vertex (because this vertex is a vertex of degree $d + 1$ now). The probability to hit a vertex of degree d is $\frac{A\,d+B}{n} + O\left(\frac{d^2}{n^2}\right)$.

Summing over all vertices of degree d we obtain that $\mathsf{E}T_n(d)$ is decreased by:

$$\left(\frac{Ad+B}{n}+O\left(\frac{d^2}{n^2}\right)\right)\cdot \mathsf{E}T_n(d). \tag{8}$$

2. Exactly one edge hits a vertex of degree $d-1$. Then $T_n(d)$ is increased by the number of triangles on this vertex. The probability to hit a vertex of degree $d-1$ once is equal to $\frac{A\,(d-1)+B}{n}+O\left(\frac{d^2}{n^2}\right)$. Summing over all vertices of degree $d-1$ we obtain that the value $\mathsf{E}T_n(d)$ is increased by:

$$\left(\frac{A(d-1)+B}{n}+O\left(\frac{d^2}{n^2}\right)\right)\cdot \mathsf{E}T_n(d-1). \tag{9}$$

3. Exactly one edge hits a vertex of degree $d-1$ and another edge hits its neighbor. Then, in addition to (9), $T_n(d)$ is increased by 1. The probability to hit a vertex of degree $d-1$ and its neighbor is equal to $\frac{D}{mn}+O\left(\frac{(d-1)\,d_i}{n^2}\right)$, where d_i is the degree of this neighbor. Summing over the neighbors of a given vertex of degree $d-1$ and summing then over all vertices of degree $d-1$ we obtain that $\mathsf{E}T_n(d)$ is increased by:

$$\begin{aligned}(d-1)\,\mathsf{E}N_n(d-1)\,\frac{D}{mn}+O\left(\frac{d\cdot \mathsf{E}\sum\limits_{\substack{i:i \text{ is a neighbor}\\ \text{of a vertex of degree } d-1}} d_i}{n^2}\right)\\ =(d-1)\,\mathsf{E}N_n(d-1)\,\frac{D}{mn}+O\left(\frac{d\,\mathsf{E}S(n,d)}{n^2}\right).\end{aligned} \tag{10}$$

4. Exactly i edges hit a vertex of degree $d-i$, where i is between 2 and m. If no edges hit the neighbors of this vertex, then $T_n(d)$ is increased only by the number of triangles on this vertex. The probability to hit a vertex of degree $d-i$ exactly i times is equal to $O\left(\frac{d^2}{n^2}\right)$. If we also hit its neighbors, then $T_n(d)$ is additionally increased by 1 for each neighbor. The probability to hit a vertex of degree $d-i$ exactly i times and hit some its neighbor is, obviously, $O\left(\frac{d^2}{n^2}\right)$. Summing over all vertices of degree $d-i$ and then summing over all i from 2 to m, we obtain that $\mathsf{E}T_n(d)$ is increased by:

$$\begin{aligned}\sum_{i=2}^{m}\left(\mathsf{E}T_n(d-i)\cdot O\left(\frac{d^2}{n^2}\right)+O\left(\frac{d^2}{n^2}\right)\cdot(d-i)\cdot \mathsf{E}N_n(d-i)\right)\\ =O\left(\frac{d^2}{n^2}\right)\mathsf{E}T_n(d)+O\left(\frac{d^3}{n^2}\right)\mathsf{E}N_n(d).\end{aligned} \tag{11}$$

Finally, using (8)–(11) and the linearity of the expectation, we get

$$\begin{aligned}
\mathsf{E}T_{n+1}(d) &= \mathsf{E}T_n(d) - \left(\frac{Ad+B}{n} + O\left(\frac{d^2}{n^2}\right)\right)\mathsf{E}T_n(d) \\
&\quad + \left(\frac{A(d-1)+B}{n} + O\left(\frac{d^2}{n^2}\right)\right)\mathsf{E}T_n(d-1) + (d-1)\,\mathsf{E}N_n(d-1)\,\frac{D}{mn} \\
&\quad + O\left(\frac{d\,\mathsf{E}S(n,d)}{n^2}\right) + O\left(\frac{d^2}{n^2}\right)\mathsf{E}T_n(d) + O\left(\frac{d^3}{n^2}\right)\mathsf{E}N_n(d) \\
&= \left(1 - \frac{Ad+B}{n}\right)\mathsf{E}T_n(d) + \frac{A(d-1)+B}{n}\,\mathsf{E}T_n(d-1) \\
&\quad + O\left(\frac{d^2}{n^2}\right)\left(\mathsf{E}T_n(d) + \mathsf{E}T_n(d-1)\right) + O\left(\frac{d^3}{n^2}\right)\mathsf{E}N_n(d) \\
&\quad + \frac{D}{mn}\,(d-1)\,\mathsf{E}N_n(d-1) + O\left(\frac{d\cdot\mathsf{E}S(n,d)}{n^2}\right). \qquad (12)
\end{aligned}$$

Consider the case $2A < 1$ (the cases $2A = 1$ and $2A > 1$ will be analyzed similarly). We prove by induction on d and n that

$$\mathsf{E}T_n(d) = K(d)\left(n + \theta\left(C \cdot d^{2+\frac{1}{A}}\right)\right) \qquad (13)$$

for some constant $C > 0$. Let us assume that $\mathsf{E}T_i(\tilde{d}) = K(\tilde{d})\left(i + \theta\left(C \cdot \tilde{d}^{2+\frac{1}{A}}\right)\right)$ for $\tilde{d} < d$ and all i and for $\tilde{d} = d$ and $i < n+1$.

Recall that $K(d) = c(m,d)\left(D + \frac{D}{m}\cdot\sum_{i=m}^{d-1}\frac{i}{Ai+B}\right)$ and $\mathsf{E}N_n(d) = c(m,d)\cdot\left(n + O\left(d^{2+\frac{1}{A}}\right)\right)$. If $2A < 1$, then from (7) and Theorem 7 we get $\mathsf{E}S(n,d) = O(n)$ and we obtain:

$$\begin{aligned}
\mathsf{E}T_{n+1}(d) &= \left(1 - \frac{Ad+B}{n}\right)K(d)\left(n + \theta\left(Cd^{2+\frac{1}{A}}\right)\right) \\
&\quad + \frac{A(d-1)+B}{n}\,K(d-1)\left(n + \theta\left(C(d-1)^{2+\frac{1}{A}}\right)\right) \\
&\quad + O\left(\frac{d^2}{n^2}\right)\left(K(d)\left(n + \theta\left(Cd^{2+\frac{1}{A}}\right)\right) + K(d-1)\left(n + \theta\left(C(d-1)^{2+\frac{1}{A}}\right)\right)\right) \\
&\quad + O\left(\frac{d^3}{n^2}\right)c(m,d)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right) \\
&\quad + \frac{D}{mn}\,(d-1)\,c(m,d-1)\left(n + O\left(d^{2+\frac{1}{A}}\right)\right) + O\left(\frac{d}{n}\right).
\end{aligned}$$

Note that $K(d) = \frac{A(d-1)+B}{Ad+B+1}\,K(d-1) + \frac{D(d-1)}{m(Ad+B+1)}\,c(m,d-1)$. Therefore, we obtain:

$$\begin{aligned}\mathsf{E}T_{n+1}(d) &= K(d)\,(n+1) + K(d)\left(1-\frac{Ad+B}{n}\right)\theta\left(C\,d^{2+\frac{1}{A}}\right)\\ &\quad + K(d-1)\,\frac{A(d-1)+B}{n}\,\theta\left(C\,(d-1)^{2+\frac{1}{A}}\right)\\ &\quad + \frac{D(d-1)}{mn}\,c(m,d)\,O\left(d^{2+\frac{1}{A}}\right) + O\left(\frac{d}{n}\right) + O\left(\frac{d^2}{n^2}\right)\,(K(d)\,n\\ &\quad + K(d)\,\theta\left(C\,d^{2+\frac{1}{A}}\right) + K(d-1)\,n + K(d-1)\,\theta\left(C\,(d-1)^{2+\frac{1}{A}}\right)\Big)\\ &\quad + O\left(\frac{d^3}{n^2}\right)\,\left(c(m,d)\,n + c(m,d)\,O\left(d^{2+\frac{1}{A}}\right)\right).\end{aligned}$$

In order to show (13), it remains to prove that for some large enough C:

$$\begin{aligned}K(d)\left(\frac{Ad+B}{n}\right)C\,d^{2+\frac{1}{A}} &\geq K(d-1)\,\frac{A(d-1)+B}{n}\,C\,(d-1)^{2+\frac{1}{A}}\\ &\quad + O\left(\frac{d^2}{n}\right) + O\left(C\,\frac{d^4}{n^2}\right) + O\left(\frac{d^4}{n^2}\right). \qquad (14)\end{aligned}$$

First, we analyze the following difference:

$$\begin{aligned}&K(d)\left(\frac{Ad+B}{n}\right)d^{2+\frac{1}{A}} - K(d-1)\,\frac{A(d-1)+B}{n}\,(d-1)^{2+\frac{1}{A}}\\ &\quad = \frac{Ad+B}{n}\,d^{2+\frac{1}{A}}\left(\frac{A(d-1)+B}{Ad+B+1}\,K(d-1) + \frac{D(d-1)}{m(Ad+B+1)}\,c(m,d-1)\right)\\ &\quad\quad - \frac{A(d-1)+B}{n}\,K(d-1)\,(d-1)^{2+\frac{1}{A}} = \frac{(Ad+B)D(d-1)}{mn(Ad+B+1)}\,c(m,d-1)\,d^{2+\frac{1}{A}}\\ &\quad\quad + K(d-1)\,\frac{A(d-1)+B}{n}\left(\frac{Ad+B}{Ad+B+1}d^{2+\frac{1}{A}} - (d-1)^{2+\frac{1}{A}}\right)\\ &\quad \geq \frac{(Ad+B)D(d-1)}{mn(Ad+B+1)}\,c(m,d-1)\,d^{2+\frac{1}{A}}\\ &\quad\quad + (d-1)^{2+\frac{1}{A}}\,K(d-1)\,\frac{A(d-1)+B}{n}\cdot\frac{2A^2d+2AB+B}{Ad(Ad+B+1)}\\ &\quad \geq \frac{(Ad+B)D(d-1)}{mn(Ad+B+1)}\,c(m,d-1)\,d^{2+\frac{1}{A}}.\end{aligned}$$

Therefore, Eq. (14) becomes:

$$C\,\frac{(Ad+B)D(d-1)}{mn(Ad+B+1)}\,c(m,d-1)\,d^{2+\frac{1}{A}} \geq O\left(\frac{d^2}{n}\right) + O\left(C\,\frac{d^4}{n^2}\right) + O\left(\frac{d^4}{n^2}\right).$$

In the case $2A = 1$ this inequality will be:

$$\begin{aligned}&C\,\frac{(Ad+B)D(d-1)}{mn(Ad+B+1)}\,c(m,d-1)\,d^{2+\frac{1}{A}}\,\log(n)\\ &\quad \geq O\left(\frac{d^2}{n}\right) + O\left(C\,\frac{d^4\cdot\log(n)}{n^2}\right) + O\left(\frac{d^4}{n^2}\right) + O\left(\frac{d\,\log(n)}{n}\right).\end{aligned}$$

In the case $2A > 1$ this inequality will be:

$$C\,\frac{(Ad+B)D(d-1)}{mn(Ad+B+1)}\,c(m,d-1)\,d^{2+\frac{1}{A}}\,n^{2A-1}$$
$$\geq O\left(\frac{d^2}{n}\right)+O\left(C\,\frac{d^4\,n^{2A-1}}{n^2}\right)+O\left(\frac{d^4}{n^2}\right)+O\left(\frac{d\,n^{2A}}{n^2}\right).$$

It is easy to see that for $n \geq Q \cdot d^2$ (for some large Q which depends only on the parameters of the model) these three inequalities are satisfied. This concludes the proof of the theorem.

4.2 Proof of Theorem 5

This theorem is proved similarly to the concentration theorem from [19]. We also need the following notation (introduced in [19]):

$$p_n(d) = \mathsf{P}\left(d_v^{n+1} = d \mid d_v^n = d\right) = 1 - A\frac{d}{n} - B\frac{1}{n} + O\left(\frac{d^2}{n^2}\right),$$
$$p_n^1(d) := \mathsf{P}\left(d_v^{n+1} = d+1 \mid d_v^n = d\right) = A\frac{d}{n} + B\frac{1}{n} + O\left(\frac{d^2}{n^2}\right),$$
$$p_n^j(d) := \mathsf{P}\left(d_v^{n+1} = d+j \mid d_v^n = d\right) = O\left(\frac{d^2}{n^2}\right),\ 2 \leq j \leq m,$$
$$p_n := \sum_{k=1}^{m} \mathsf{P}(d_{n+1}^{n+1} = m+k) = O\left(\frac{1}{n}\right).$$

To prove Theorem 5 we also need the Azuma–Hoeffding inequality:

Theorem 8 (Azuma, Hoeffding). *Let $(X_i)_{i=0}^n$ be a martingale such that $|X_i - X_{i-1}| \leq c_i$ for any $1 \leq i \leq n$. Then $\mathsf{P}\left(|X_n - X_0| \geq x\right) \leq 2e^{-\frac{x^2}{2\sum_{i=1}^n c_i^2}}$ for any $x > 0$.*

Consider the random variables $X_i(d) = \mathsf{E}(T_n(d) \mid G_m^i)$, $i = 0, \ldots, n$. Note that $X_0(d) = \mathsf{E}T_n(d)$ and $X_n(d) = T_n(d)$. It is easy to see that $X_n(d)$ is a martingale.

We will prove below that for any $i = 0, \ldots, n-1$

(1) if $2A < 1$, then $|X_{i+1}(d) - X_i(d)| \leq Md^2$,
(2) if $2A = 1$, then $|X_{i+1}(d) - X_i(d)| \leq Md^2 \log(n)$,
(3) if $1 < 2A < \frac{3}{2}$, then $|X_{i+1}(d) - X_i(d)| \leq Md^2 n^{2A-1}$,

where $M > 0$ is some constant. The theorem follows from this statement immediately. Indeed, consider the case $2A < 1$. Put $c_i = Md^2$ for all i. Then from Azuma–Hoeffding inequality it follows that

$$\mathsf{P}\left(|T_n(d) - \mathsf{E}T_n(d)| \geq d^2\sqrt{n}\,\log n\right) \leq 2\exp\left\{-\frac{n\,d^4\,\log^2 n}{2\,n\,M^2 d^4}\right\} = O\left(n^{-\log n}\right).$$

Therefore, for the case $2A < 1$ the first statement of the theorem is satisfied. If $d \le n^{\frac{A-\delta}{4A+2}}$, then the value $n\,d^{-1/A}$ is considerably greater than $d^2 \log n \sqrt{n}$. From this the second statement of the theorem follows. The cases $2A = 1$ and $2A > 1$ can be considered similarly. It remains to estimate $|X_{i+1}(d) - X_i(d)|$.

Fix $0 \le i \le n-1$ and some graph G_m^i. Note that

$$\left|\mathsf{E}\left(T_n(d) \mid G_m^{i+1}\right) - \mathsf{E}\left(T_n(d) \mid G_m^i\right)\right| \le \max_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathsf{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\} - \min_{\tilde{G}_m^{i+1} \supset G_m^i} \left\{\mathsf{E}\left(T_n(d) \mid \tilde{G}_m^{i+1}\right)\right\}.$$

Put $\hat{G}_m^{i+1} = \arg\max \mathsf{E}(T_n(d) \mid \tilde{G}_m^{i+1})$, $\bar{G}_m^{i+1} = \arg\min \mathsf{E}(T_n(d) \mid \tilde{G}_m^{i+1})$. It is sufficient to estimate the difference $\mathsf{E}(T_n(d) \mid \hat{G}_m^{i+1}) - \mathsf{E}(T_n(d) \mid \bar{G}_m^{i+1})$.

For $i+1 \le t \le n$ put

$$\delta_t^i(d) = \mathsf{E}(T_t(d) \mid \hat{G}_m^{i+1}) - \mathsf{E}(T_t(d) \mid \bar{G}_m^{i+1}).$$

First, let us note that for $n \le W \cdot d^2$ (the value of constant W will be defined later) we have $\delta_n^i(d) \le \frac{2mn}{d} \cdot \left(\frac{m(m-1)}{2} + d\,m\right) \le 4m^2 n \le Md^2 \le Md^2 \log(n) \le Md^2 n^{2A-1}$ (since we have at most $\frac{2mn}{d}$ vertices of degree d, and each vertex of degree d has at most $\frac{m(m-1)}{2}$ triangles when this vertex appears plus at each step we get a triangle only if we hit both the vertex under consideration and a neighbor of this vertex, and our vertex degree is equal to d, therefore we get at most $d\,m$ triangles) for some constant M which depends only on W and m.

It remains to estimate $\delta_n^i(d)$ for $n > Wd^2$. Consider the case $2A < 1$. We want to prove that $\delta_n^i(d) \le Md^2$ for $n > Wd^2$ by induction. Suppose that $n = i+1$. Fix G_m^i. Graphs $\hat{G}_m^{i+1}$ and $\bar{G}_m^{i+1}$ are obtained from the graph G_m^i by adding the vertex $i+1$ and m edges. These m edges can affect the number of triangles on at most m previous vertices. For example, they can be drown to at most m vertices of degree d and decrease $T_i(d)$ by at most $\frac{m\,d\,(d-1)}{2}$. Such reasonings finally lead to the estimate $\delta_{i+1}^i(d) \le Md^2$ for some M.

Now let us use the induction. Consider t: $i+1 \le t \le n-1$, $t > W\,d^2$ (note that the smaller values of t were already considered). Using similar reasonings as in the proof of Theorem 4 we get:

$$\delta_{t+1}^i(m) = \delta_t^i(m)\,(1 - p_t(m)) + O\left(\frac{1}{t}\right),$$

$$\begin{aligned}\delta_{t+1}^i(d) = {} & \delta_t^i(d)\,(1 - p_t(d)) + \delta_t^i(d-1)\,p_t^1(d-1) \\ & + (d-1) \cdot \left(\mathsf{E}(N_t(d-1) \mid \hat{G}_m^i) - \mathsf{E}(N_t(d-1) \mid \bar{G}_m^i)\right) \cdot \frac{D}{mt} \\ & + O\left(\frac{d \cdot \mathsf{E}S(t, d-1)}{t^2}\right) + O\left(\frac{\mathsf{E}T_t(d) \cdot d^2}{t^2}\right) + O\left(\frac{\mathsf{E}N_t(d) \cdot d^3}{t^2}\right).\end{aligned}$$

Note that $\mathsf{E}(N_t(d) \mid \hat{G}_m^{i+1}) - \mathsf{E}(N_t(d) \mid \bar{G}_m^{i+1}) = O(d)$ (see [19]) and $\mathsf{E}S(t, d-1) = O(t)$. From this recurrent relations it is easy to obtain by induction that $\delta_n^i(d) \le Md^2$ for some M. Indeed,

$$\delta_{t+1}^i(m) \le Mm^2(1 - p_t(m)) + \frac{C_1}{t} \le Mm^2\left(1 - \frac{Am+B}{t} + \frac{C_2}{t^2}\right) + \frac{C_1}{t} \le Mm^2$$

for sufficiently large M. By C_i, $i = 1, 2, \ldots$, we denote some positive constants. For $d > m$ we get

$$\begin{aligned}
\delta_{t+1}^i(d) &\le Md^2(1 - p_t(d)) + M(d-1)^2 p_t^1(d-1) + C_3\frac{d^2}{t} + C_4\frac{d^4}{t^2} \\
&\le Md^2\left(1 - \frac{Ad+B}{t} + C_5\frac{d^2}{t^2}\right) + M(d-1)^2\left(\frac{A(d-1)+B}{t} + C_6\frac{d^2}{t^2}\right) + C_3\frac{d^2}{t} \\
&\quad + C_4\frac{d^4}{t^2} \le Md^2 + \frac{M}{t}\left(A(-3d^2+3d-1) + B(-2d+1) + C_7\frac{d^4}{t} + C_3\frac{d^2}{M} + C_4\frac{d^4}{Mt}\right) \\
&\le Md^2 + \frac{M}{t}\left(\left(-3A + C_7\frac{d^2}{t} + \frac{C_3}{M} + C_4\frac{d^2}{Mt}\right)\cdot d^2\right. \\
&\quad \left. + (3A - 2B)\cdot d + (B - A)\right) \le Md^2.
\end{aligned}$$

for sufficiently large W and M.

In the case $2A = 1$ we have $\mathsf{E}S(t, d-1) = O(t\log(t))$ and we get the following inequalities:

$$\delta_{t+1}^i(m) \le Mm^2\log(t)(1 - p_t(m)) + \frac{C_1\log(t)}{t} \le Mm^2\log(t+1),$$

$$\begin{aligned}
\delta_{t+1}^i(d) &\le Md^2\log(t)(1 - p_t(d)) + M(d-1)^2\log(t)\,p_t^1(d-1) \\
&\quad + C_2\frac{d^2}{t} + C_3\frac{d\log(t)}{t} + C_4\frac{d^4\log(t)}{t^2} \le Md^2\log(t+1).
\end{aligned}$$

In the case $2A > 1$ we have $\mathsf{E}S(t, d-1) = O\left(t^{2A}\right)$ and we get the following inequalities:

$$\delta_{t+1}^i(m) \le Mm^2t^{2A-1}(1 - p_t(m)) + \frac{C_1t^{2A-1}}{t} \le Mm^2(t+1)^{2A-1},$$

$$\begin{aligned}
\delta_{t+1}^i(d) &\le Md^2t^{2A-1}(1 - p_t(d)) + M(d-1)^2\,t^{2A-1}p_t^1(d-1) \\
&\quad + C_2\frac{d^2}{t} + C_3\frac{d\cdot t^{2A-1}}{t} + C_4\frac{d^4t^{2A-1}}{t^2} \le Md^2(t+1)^{2A-1}.
\end{aligned}$$

This concludes the proof of Theorem 5.

5 Conclusion

In this paper, we study the local clustering coefficient $C(d)$ for the vertices of degree d in the T-subclass of the PA-class of models. Despite the fact that the T-subclass generalizes many different models, we are able to analyze the local clustering coefficient for all these models. Namely, we proved that $C(d)$ asymptotically decreases as $\frac{2D}{Am} \cdot d^{-1}$. In particular, this result implies that one cannot change the exponent -1 by varying the parameters A, D, and m. This basically means that preferential attachment models in general are not flexible enough to model $C(d) \sim d^{-\psi}$ with $\psi \neq 1$.

We would also like to mention the connection between the obtained result and the notion of *weak* and *strong transitivity* introduced in [21]. It was shown in [22] that percolation properties of a network are defined by the type (weak or strong) of its connectivity. Interestingly, a model from the T-subclass can belong to either weak or strong transitivity class: if $2D < Am$, then we obtain the weak transitivity; if $2D > Am$, then we obtain the strong transitivity.

References

1. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
2. Bansal, S., Khandelwal, S., Meyers, L.A.: Exploring biological network structure with clustered random networks. BMC Bioinf. **10**, 405 (2009)
3. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Sci. **286**(5439), 509–512 (1999)
4. Barabási, A.-L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Phys. A **272**(1–2), 173–187 (1999)
5. Albert, R., Jeong, H., Barabási, A.-L.: Internet: diameter of the world-wide web. Nat. **401**, 130–131 (1999)
6. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: structure and dynamics. Phys. Rep. **424**(45), 175–308 (2006)
7. Bollobás, B., Riordan, O.M.: Mathematical results on scale-free random graphs. In: Handbook of Graphs and Networks: From the Genome to the Internet (2003)
8. Bollobás, B., Riordan, O.M., Spencer, J., Tusnády, G.: The degree sequence of a scale-free random graph process. Random Struct. Algorithms **18**(3), 279–290 (2001)
9. Borgs, C., Brautbar, M., Chayes, J., Khanna, S., Lucier, B.: The power of local information in social networks. Preprint (2012)
10. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Comput. Netw. **33**(16), 309–320 (2000)
11. Buckley, P.G., Osthus, D.: Popularity based random graph models leading to a scale-free degree sequence. Discrete Math. **282**, 53–63 (2004)
12. Catanzaro, M., Caldarelli, G., Pietronero, L.: Assortative model for social networks. Phys. Rev. E **70**, 037101 (2004)
13. Leskovec, J.: Dynamics of large networks. ProQuest (2008)
14. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: Proceedings of SIGCOMM (1999)

15. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Nat. Acad. Sci. **99**(12), 7821–7826 (2002)
16. Holme, P., Kim, B.J.: Growing scale-free networks with tunable clustering. Phys. Rev. E **65**(2), 026107 (2002)
17. Newman, M.E.J.: Pareto distributions and Zipf's law. Contemp. Phys. **46**(5), 323–351 (2005)
18. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
19. Ostroumova, L., Ryabchenko, A., Samosvat, E.: Generalized preferential attachment: tunable power-law degree distribution and clustering coefficient. In: Bonato, A., Mitzenmacher, M., Prałat, P. (eds.) WAW 2013. LNCS, vol. 8305, pp. 185–202. Springer, Heidelberg (2013)
20. Ravasz, E., Barabási, A.-L.: Hierarchical organization in complex networks. Phys. Rev. E **67**(2), 26112 (2003)
21. Serrano, M.A., Boguñá, M.: Clustering in complex networks. I. General formalism. Phys. Rev. E **74**, 056114 (2006)
22. Serrano, M.A., Boguñá, M.: Clustering in complex networks. II. Percolation properties. Phys. Rev. E **74**, 056115 (2006)
23. Vázquez, A., Pastor-Satorras, R., Vespignani, A.: Large-scale topological and dynamical properties of the internet. Phys. Rev. E **65**, 066130 (2002)
24. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)
25. Zhou, T., Yan, G., Wang, B.-H.: Maximal planar networks with large clustering coefficient and power-law degree distribution. Phys. Rev. E **71**(4), 046141 (2005)

http://www.springer.com/978-3-319-26783-8

Algorithms and Models for the Web Graph
12th International Workshop, WAW 2015, Eindhoven,
The Netherlands, December 10-11, 2015, Proceedings
Gleich, D.F.; Komjathy, J.; Litvak, N. (Eds.)
2015, VIII, 203 p. 17 illus. in color., Softcover
ISBN: 978-3-319-26783-8