3

Parametric network generation model: Context dependent preferential attachment

In this chapter we propose a growing random network model for generation of complex networks which exhibit power-law in the tail of its degree distribution. Several such models in literature start with an initial small network and a new node appears at each time step, and gets linked with one or multiple existing nodes. One of the fundamental models in this direction was first proposed by Albert and Barabasi in [Barabási and Albert, 1999a] that is called BA model. In this model, the new node prefers to get connected with an existing node which has high degree.

In this context, the following question can be raised. Does a new node always wish to form links with important (high degree) nodes or the choice get influenced by other factors also? Moreover, if the choice gets influenced by other properties of the existing nodes, will the network be having power-law degree distribution? An evidence of a phenomena that people's choice does not depend on only one property is given in [Tversky and Simonson, 1993] supported by an empirical data (see [Tversky, 1972; Huber *et al.*, 1982; Simonson, 1989] also). The data shows that at the time of purchasing a product, a buyer considers the background (history) of the product and relative attractiveness of the product with respect to other products in the same reference. Thus, the concept of context preferential attachment was introduced in [Tversky and Simonson, 1993].

We propose a growing random complex network model where the probability of link formation is determined by weighted local and global property of the existing nodes. We consider that local and global properties of a node as degree and the relative average degree of the node in the existing network. Thus, we call the proposed model, the context dependent preferential attachment model (CDPAM) for complex networks. We prove that the degree distribution of complex networks generated by CDPAM follow power law $P(k) = L(k)k^{-\gamma}$ where $2 \leq \gamma \leq 3$ and $L(k) \to \alpha$ (a constant which depends on the weights given on local and global property of the nodes) as $k \to \infty$. We also prove that the expected diameter grows logarithmically with the size of the new nodes added in the network, however the growth of the expected diameter is slower than that of the BA model. The numerical simulations show that the expected diameter stabilizes when alike weights are given to the local and global property which determine the preference of link formation. In contrast to the conventional wisdom that diameter shows as a function of $\ln(\ln N)$ or $\ln N$ in real networks, the authors in [Leskovec *et al.*, 2007] observed that the diameter stabilizes or shrinks as a network grows. The proposed model reveals how shrinking and increasing of diameter are related to the weights on local and global property of the nodes during expansion of the network.

A variety of mathematical and statistical measures have been proposed in literature in order to characterize global and local structure of complex networks. We derive clustering coefficient, assortativity, number of triangles, algebraic connectivity and spectral radius for different complex networks generated by CDPAM and compare them with the same obtained from the complex networks generated by BA model. We show that CDPAM replicates properties of real networks for all these measures when alike weights are given to local and global property. Finally, we observe that the BA model is a limiting case of CDPAM when new nodes tend to give large weight to the local property compared to the weight given on the global property during link formation.

3.1 CONTEXT DEPENDENT PREFERENTIAL ATTACHMENT MODEL (CDPAM)

In this section, we propose a random complex network model which relies on the fact that the network is open i.e. a network continuously grows in time with the addition of new nodes in to a fixed small network chosen in the beginning of the process [Barabási *et al.*, 1999]. It is important to notice that the link formation in BA model is biased as the link formation depends only on the high degree (importance) of the existing nodes. However, in real life individuals prefer to form relationship (link) with important (global property) people in society but also give importance to background (local property) of the people before making the relation. Inspired by this thought, we introduce the model as follows.

1. Growth: Starting with a small network having m_0 nodes, at every timestep add a new node with $m \leq m_0$ edges such that degree of any node in the initial network should lie between m to $2m$.

2. Context preferential attachment: Assume that $V(t)$ denotes the node set of the network after t-time step. When a new node j appears at time $t+1$, it will get connected to a node $i \in V(t)$ with probability $p_j^i(t+1)$ given by

$$p_j^i(t+1) = \frac{\beta f_B(i) + \theta g(i, V(t))}{\sum_{i \in V(t)} (\beta f_B(i) + \theta g(i, V(t)))} \tag{3.1}$$

   where $f_B(i)$ quantifies the background (local context) of node i, $g(i, V(t))$ determines the relative advantage (global context) of a nodes over others in the network $G(t)$, and $\beta, \theta (< \beta)$ are the positive control parameters for the property of the nodes in $V(t)$.

   In order to simplify the model, we consider

$$f_B(i) = k_i \, \text{and} \, g(i, V(t)) = \frac{\sum_{l \in V(t)} k_i - k_l}{|V(t)|}$$

where k_i denotes the degree of a node i and $|V(t)|$ is the number of nodes in $G(t)$. As we consider that a single node appears at each timestep, after time t there will be $t + m_0$ nodes in the network and for a large value of $t (\gg m_0)$, $|V(t)| \approx t$. Consequently, we have

$$p_j^i(t+1) \approx \frac{\beta k_i + \theta \sum_{l \in V(t)} \frac{k_i - k_l}{t}}{\sum_{l \in V(t)} \beta k_l + \sum_{l \in V(t)} \frac{(t+m_0)k_l - 2mt - m_0(m_0-1)}{t}} \approx \frac{\beta k_i + \theta (k_i - 2m)}{2m\beta t}$$

for a very small value of m_0. Assuming k_i to be a continuous real variable function and the rate of change of k_i is proportional to $p_i^j(t)$, we have

$$\frac{\partial k_i}{\partial t} = m \frac{\beta k_i + \theta (k_i - 2m)}{2m\beta t} \tag{3.2}$$

by applying mean field theory.

Further, by (3.2),

$$\frac{\partial k_i}{\partial t} = m \frac{\beta k_i + \theta (k_i - 2m)}{2m\beta t} = \frac{k_i - c}{(\gamma - 1)t}, \quad \gamma = 1 + \frac{2\beta}{\beta + \theta}, c = \frac{2m\theta}{\beta + \theta}$$

solving which we obtain

$$k_i(t) = (m-c)\left(\frac{t}{t_i}\right)^{1/(\gamma-1)} + c \tag{3.3}$$

when the initial condition is given by $k_i(t_0) = m$. This yields

$$P(k_i(t) < k) = P(t_i > (m-c)^{\gamma-1}(k-c)^{1-\gamma}t).$$

Assuming $k_i(t) < k$, we have $t_i > (m-c)^{\gamma-1}(k-c)^{1-\gamma}t$, where t_i is the time when i^{th} node appears in the network . Further, since it is assumed that a single node gets added at each timestep, it is equivalent to a uniform distribution of t_i, given by $P(t_i) = 1/(m_0+t)$. Consequently,

$$\begin{aligned} P(k_i(t) < k) &= P(t_i > (m-c)^{\gamma-1}(k-c)^{1-\gamma}t) \\ &= 1 - \tfrac{t}{t+m_0}(m-c)^{\gamma-1}(k-c)^{1-\gamma}. \end{aligned}$$

Thus, the degree distribution is given by

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{t}{t+m_0}(\gamma-1)(m-c)^{\gamma-1}(k-c)^{-\gamma} = L(k)k^{-\gamma}$$

where $L(k) \to (\gamma-1)(m-c)^{(\gamma-1)}$ as $k \to \infty$. In particular, $\gamma \approx 2$ if $\beta \approx \theta$ and $\gamma \approx 3$ if $\beta \gg \theta$.

Setting the initial network as the complete network (fully connected network) with 7 nodes, i.e. $m_0 = 7$ and $m = 5$, we plot degree distributions of complex networks of 10000 nodes generated by CDPAM for different values of β and γ given in Figs. 3.1(a) to 3.5(b). We also calculate the p-value which is a measure of goodness-of-fit based on KS statistics, to validate the power-law degree distribution of the networks [Clauset *et al.*, 2009a]. The numerical simulations show that the exponent γ is an increasing function of β when θ is fixed.

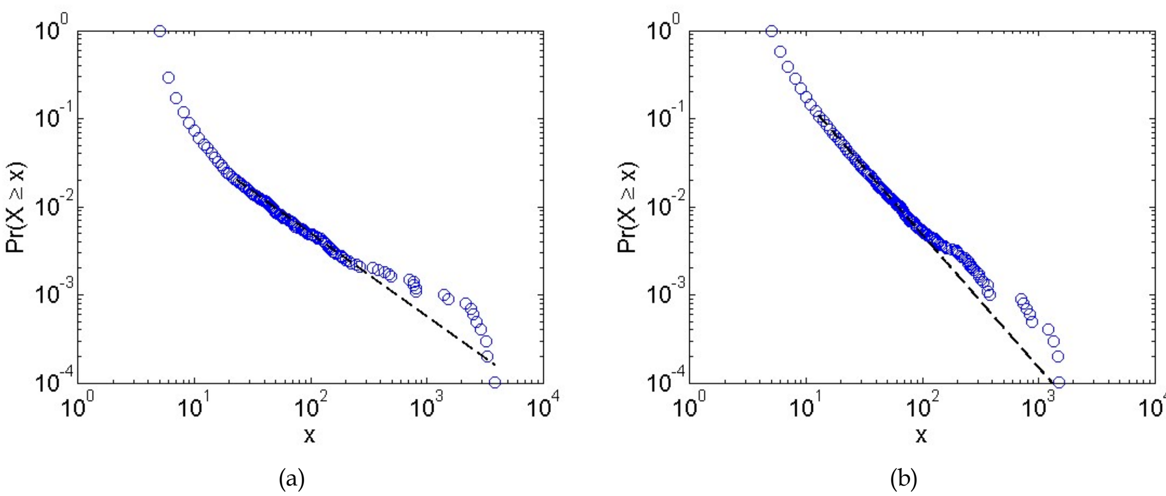
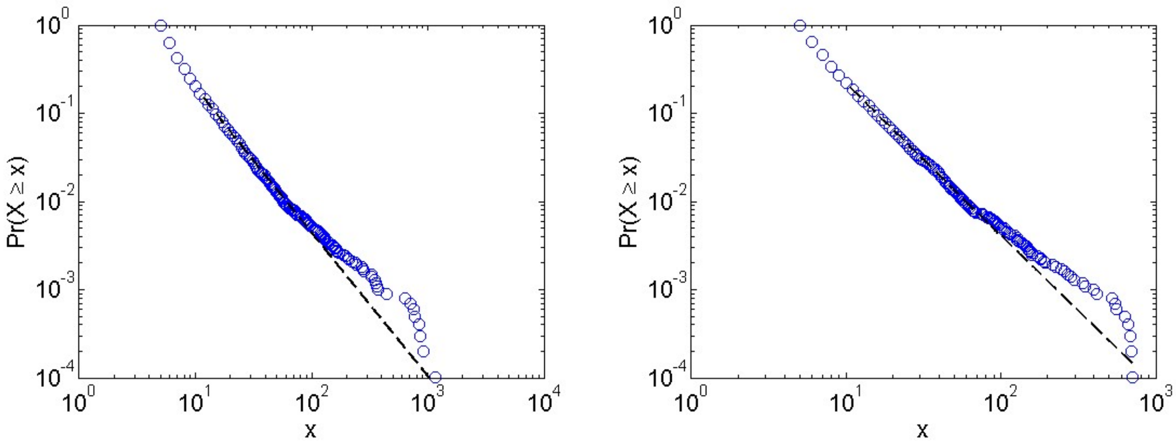


Figure 3.1 : Degree distribution of different networks of size 10000 generated by CDPAM with different values of β when $\theta = 0.5$. In (a) $\beta = 0.6$ and in (b) $\beta = 1.2$.

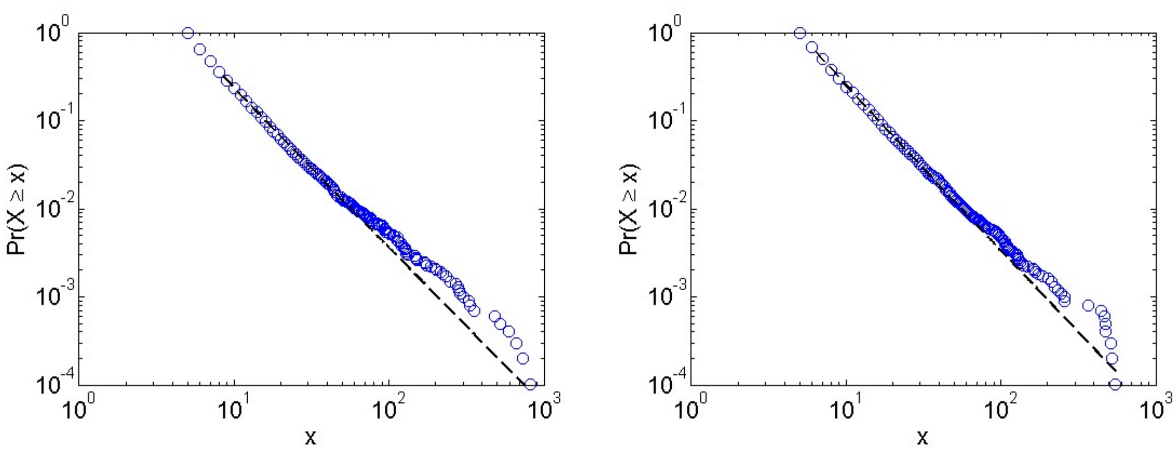
In order to show that the diameter of a complex network constructed by the CDPAM is small, we proceed as follows. Suppose the nodes i and j appeared in the network at time t_i and t_j respectively where $t_i < t_j$. Then the probability of the node j to be linked with the node i is given by

$$p^i_j = m\frac{\beta k_i(t_j) + \theta(k_i(t_j) - 2m)}{2m\beta t_j}$$



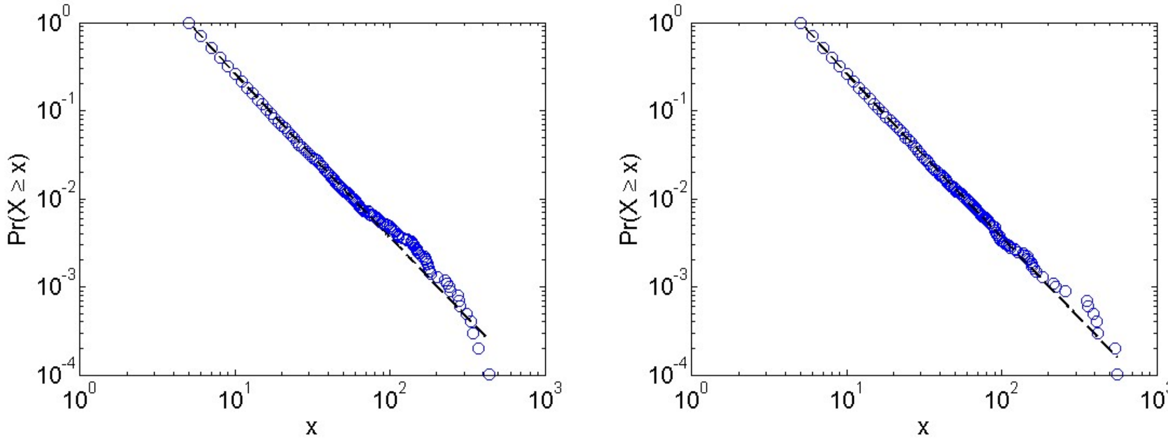
(a) (b)

Figure 3.2 : Degree distribution of different networks of size 10000 generated by CDPAM with different values of β when $\theta = 0.5$. In (a) $\beta = 1.8$ and in (b) $\beta = 2.4$.



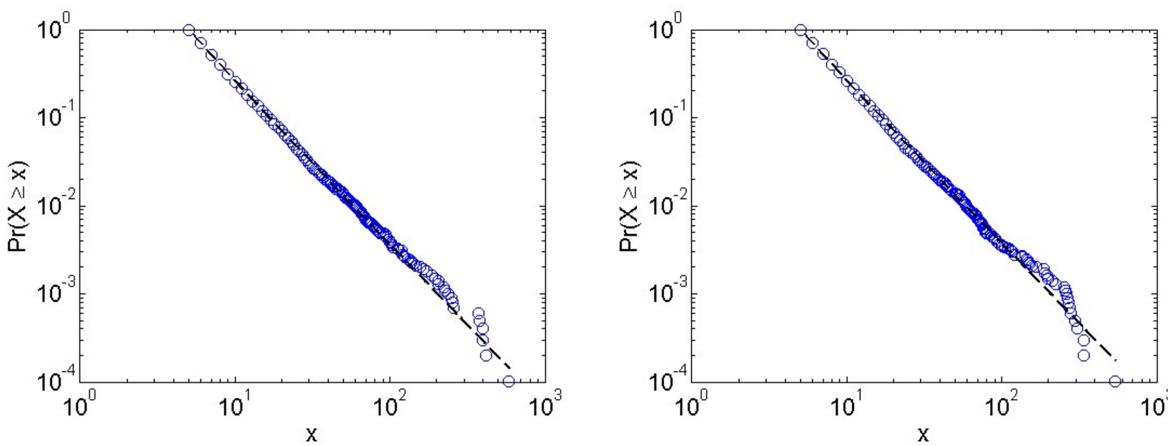
(a) (b)

Figure 3.3 : Degree distribution of different networks of size 10000 generated by CDPAM with different values of β when $\theta = 0.5$. In (a) $\beta = 3$ and in (b) $\beta = 6$.



(a) (b)

Figure 3.4 : Degree distribution of different networks of size 10000 generated by CDPAM with different values of β when $\theta = 0.5$. In (a) $\beta = 60$ and in (b) $\beta = 300$.



(a) (b)

Figure 3.5 : Degree distribution of different networks of size 10000 generated by CDPAM with different values of β when $\theta = 0.5$. In (a) $\beta = 600$ and in (b) $\beta = 600000$.

where $k_i(t_j) = (m-c)\left(\frac{t_j}{t_j}\right)^{1/(\gamma-1)} + c$ (see (3.3)) is the expected degree of the node i at time t_j. Thus,

$$p_j^i = \frac{m-c}{(\gamma-1)t_i^{1/(\gamma-1)}t_j^{1-1/(\gamma-1)}} + \frac{m(1-2\theta)}{2\beta t_j}. \tag{3.4}$$

Remark 3.1.1. *It is evident from the above derivation that the control parameters β and θ which represent weights to the local and global property of the existing nodes respectively, determine the topology of the network generated by CDPAM. A natural question would be: Does there exist a functional relation between these parameters? To investigate how different values of these parameters affect the topology of the network, we fix the parameter θ and vary β in the sequel. Thus, now onward we set $\theta = 0.5$.*

We recall the following lemma from [Fronczak *et al.*, 2004].

Lemma 3.1.2. *If $A_1, A_2, ...A_n$ are mutually independent events and their probabilities fulfil the relations $P(A_i) \leq \varepsilon$ for all i then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - exp\left(-\sum_{i=1}^{n} P(A_i)\right) - Q$$

where $0 \leq Q < \sum_{j=0}^{n+1}(n\varepsilon)^j/j! - (1+\varepsilon)^n$.

Let $V(t)$ be the set of all nodes which have been added in to the initial network up to timestep t during the growing process of the network formation. Thus, $|V(t)|$, the number of nodes in $G(t)$, is the size of new nodes added in the network formed by CDPAM up to timestep t. For any $i, j \in V(t)$, an event A_k is defined as the existence of a path of length l between i and j. The total number of such events possible is $|V(t)|^{l-1}$. Thus, as given in [Fronczak *et al.*, 2004], the probability of the existence of a path between i and j of length not more than l is given by

$$\begin{aligned} P_{ij}(l) &= P\left(\bigcup_{k=1}^{|V(t)|^{l-1}} A_k\right) \\ &= 1 - exp\left[-\sum_{v_1=1}^{|V(t)|} \cdots \sum_{v_{l-1}=1}^{|V(t)|} p_i^{v_1} \cdots p_{v_{l-1}}^{j}\right]. \end{aligned} \tag{3.5}$$

We use this result to obtain the following corollary.

Table 3.1 : Theoretically and numerically calculated values of γ of the networks of size 10000 generated by CDPAM with different values of β when $\theta = 0.5$

β ($\theta = 0.5$)	γ (*calculated Numerically*)	*p-value*	γ (*Theoretical*)
0.6	1.94	0.170	2.090
1.2	2.50	0.090	2.411
1.8	2.62	0.220	2.565
2.4	2.70	0.490	2.655
3.0	2.78	0.135	2.714
6	2.84	0.025	2.846
60	2.82	0.600	2.980
300	2.82	0.996	2.996
600	2.82	0.290	2.998
600000	2.81	0.017	2.999

Corollary 3.1.3. *The probability of the existence of a path between two vertices $i, j \in V(t)$ of length not more than l is given by*

$$P_{ij}(l) = 1 - exp\left[-\frac{K^l H_n^{l-1}}{t_i^{1/(\gamma-1)} t_j^{1-1/(\gamma-1)}}\right]$$

where $K = \frac{(\beta+0.5)(m-c)}{2\beta}, H_n = \sum_{k=1}^{|V(t)|} \frac{1}{k}$ *and* $c = \frac{m}{\beta+0.5}$.

Proof: Using (3.4) and (3.5) the result follows.

Corollary 3.1.4. *The expected value l_{ij} of the distance between two nodes $i, j \in V(t)$ is given by*

$$l_{ij} = \frac{\left(1 - \frac{1}{\gamma-1}\right) \ln t_j + \frac{1}{\gamma-1} \ln t_i + \ln H_n - r}{\ln(KH_n)} + \frac{1}{2}.$$

Proof: The result follows from the fact that

$$l_{ij} = \sum_{l=0}^{\infty} F(l)$$

where $F(l) = 1 - P_{ij}(l)$ (see [Fronczak *et al.*, 2004]).

It follows from the corollary 3.1.4 that the expected distance l_{ij} between two nodes $i, j \in V(t)$ is an increasing function of t_i and t_j when other parameters are fixed. This implies that the diameter of the network is the expected distance between the first node and the last node added in the network. Hence, setting $t_j = |V(t)|$ and $t_i = 1$ we obtain the following result.

Corollary 3.1.5. *The expected diameter of a complex network generated by CDPAM is given by*

$$D_G = \frac{\left(1 - \frac{1}{\gamma-1}\right) \ln |V(t)| + \ln H_n - r}{\ln(KH_n)} + \frac{1}{2}.$$

Thus it follows from the above corollary that the expected diameter of the network depends on the logarithmic value of the size of new nodes added in the network. In Fig. 3.6, we calculated the expected diameter for CDPAM and the approximate diameter given by BA model ($\sim \ln |V|/\ln\ln |V|$) where $|V|$ denotes the size of the entire network [Cohen and Havlin, 2003]. However, numerical simulations show that the expected diameter of CDPAM stabilizes when alike weights are assigned to both the local and global properties which determine the preference of link formation. In contrast to the conventional wisdom that diameter is a function of $\ln(\ln |V|)$ or $\ln |V|$ in real networks, the authors in [Leskovec *et al.*, 2007] observed that the diameter stabilizes or shrinks as a network grows. The CDPAM reveals how shrinking and increasing of diameter are related to the weights on local and global property of the nodes during expansion of the network.

3.2 PROPERTY OF COMPLEX NETWORKS GENERATED BY CDPAM

In this section, we numerically calculate various measures which include clustering coefficient, assortativity, algebraic connectivity, and spectral radius of the complex networks generated by CDPAM. These measures determine various topological features of a network and enable to compare how the proposed model captures the property of different real networks. We also compare values of these measures with that of complex networks generated by BA model.

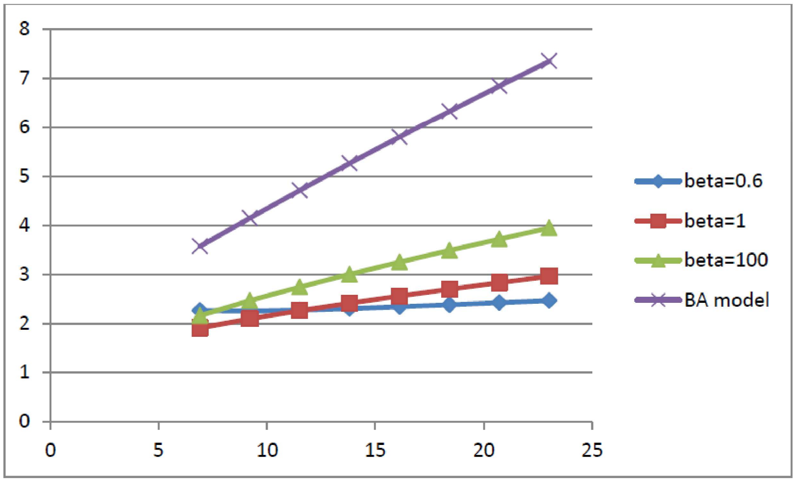


Figure 3.6 : Growth of the diameter of networks. Horizontal-axis represents the logarithm of the number of nodes ($\ln |V|$) and the vertical-axis represents diameter (D). As β decreases to $\theta = 0.5$, the growing rate of the diameter reduces.

All the measures below are calculated for networks generated by CDPAM and BA model while the initial network is considered as the fully connected network having $m_0 = 7$ nodes. In the growing process, at each timestep, one node is added to the existing network with $\bar{k} = 5$ new links made by the new node for different values of β when $\theta = 0.5$. We have used MATLAB R2012a for the numerical simulations.

3.2.1 Clustering coefficient

Clustering coefficient (CC) of a node signifies the local edge density among the neighbors of the node. The CC of a network is the average of CC of all the nodes. Thus, for a network G,

$$CC(i) = \frac{2|E_i|}{k_i(k_i - 1)} \text{ and } CC(G) = \frac{1}{|V|}\sum_i CC(i)$$

where $|E_i|$ denotes the number of links adjacent to a node i of the network [Watts and Strogatz, 1998b]. It is evident that $0 \le CC(G) \le 1$ for any network G. In Fig. 3.7, we plot the CC of different size of complex networks generated by CDPAM with different values of β and $\theta = 0.5$. It shows that as the value of β increases the CC of the network decreases and eventually when β is very large, the CC is close to zero which is a phenomena for networks generated by BA model. The Fig. 3.8 shows that the CC gets close to 0.8 as $\log \beta$ gets close to zero. Thus, we conclude that, in CDPAM model, if links are formed by giving similar weights to both the local and global properties of the existing nodes then the CC gets close to 0.8 which is a property of a large class of real networks like ego-Facebook network, ego-Gplus network, ego-Twitter network [Leskovec and Mcauley, 2012].

3.2.2 Assortativity index

The Assortative Index (AI) of a network G is defined by

$$AI(G) = \frac{\sum_{ij}(a_{ij} - \frac{k_i k_j}{2m})k_i k_k}{\sum_{ij}(k_i \delta_{ij} - \frac{k_i k_j}{2m})k_i k_j}$$

where a_{ij} is the ij^{th} entry of the adjacency matrix associated with G, δ_{ij} is the Kronecker delta function [Newman, 2002]. Obviously $-1 \le AI(G) \le 1$. A positive value of $AI(G)$ signifies nodes

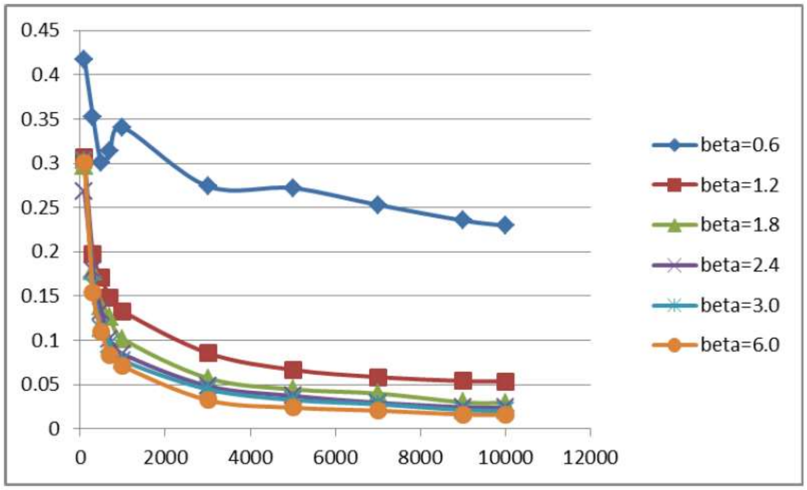


Figure 3.7 : Average clustering coefficient is plotted for different size of networks with different values of β and $\theta = 0.5$. Horizontal-axis represents the number of nodes of a network and the vertical-axis represents average clustering coefficient of the network.

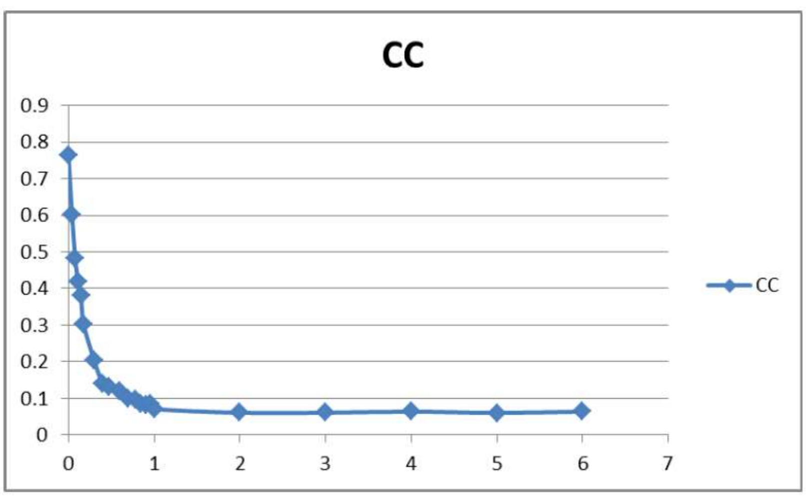


Figure 3.8 : Average clustering of the networks having 1000 nodes generated by CDPAM with different values of β when $\theta = 0.5$. Horizontal-axis represents $\log \beta$ and the vertical-axis represents average clustering coefficient of the network.

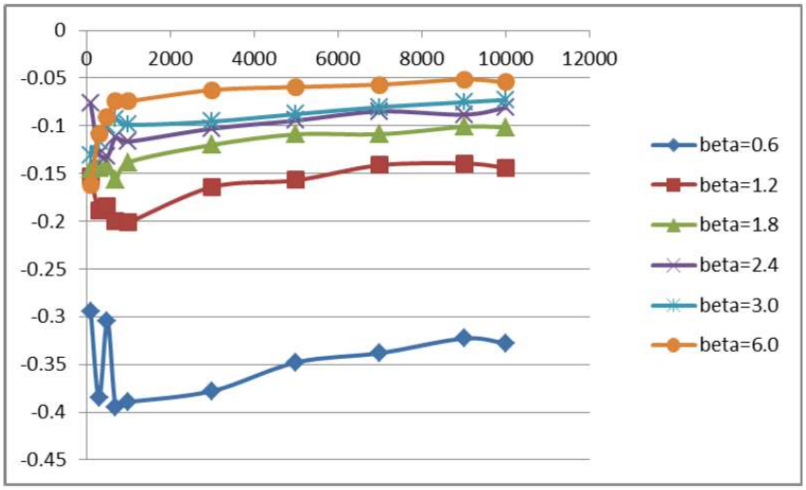


Figure 3.9 : Assortativity index is plotted for different size of networks generated by CDPAM with different values of β when $\theta = 0.5$. Horizontal-axis represents the number of nodes of the networks and the vertical-axis represents assortativity index of the networks.

with similar degrees are linked whereas a negative value of $AI(G)$ implies that similar degree nodes are not linked.

Observe that the degree of a node is a decreasing function of the timestep at which it gets added, follows by (3.3). Further, if a node j is added at the t_j timestep then the probability p_j^i to get it linked with an existing node i appeared at $t_i < t_j$ is a decreasing function of both t_i and t_j, see (3.4). These indicate, the probability of having a link between high degree nodes is larger compared to the probability of having a link in between low degree nodes. Therefore, we conclude that the network is assortative for higher degree nodes and disassortative for low degree nodes. Since the network has a few high degree nodes, overall the network shows disassortative mixing behaviour. The plots given in Fig. 3.9 assert the same for different values of β when $\theta = 0.5$. We mention that the disassortative phenomena of networks occur in a large class of real networks including World-Wide-Web [Barabási and Albert, 1999a], Marine food web [Huxham *et al.*, 1996], freshwater food web [Martinez, 1991].

3.2.3 Number of triangles

A triangle is a cycle with three nodes. The number of triangles is a fundamental building block for many real networks. In a social network, if nodes are human beings and links are described by friendship relation, then the existence of a triangle means friends of a friend are friends. Often real networks consist of a huge number of triangles for example ego-Facebook network, ego-Gplus network, ego-Twitter [Leskovec and Mcauley, 2012]. In Fig. 3.10, we show that the complex networks generated by CDPAM contain huge number of triangles compared to networks constructed by the BA model.

3.2.4 Algebraic connectivity

Algebraic connectivity of a network G is the second smallest eigenvalue of the Laplacian matrix $L = D - A$ associated with the network where $D = \text{diag}\{k_1, \ldots, k_n\}$ denotes the degree matrix and A is the adjacency matrix of the network [Fiedler, 1973]. Obviously, L is a symmetric positive semi-definite matrix. It is well known that the second eigenvalue $\lambda_2(L)$ of L is positive if and only

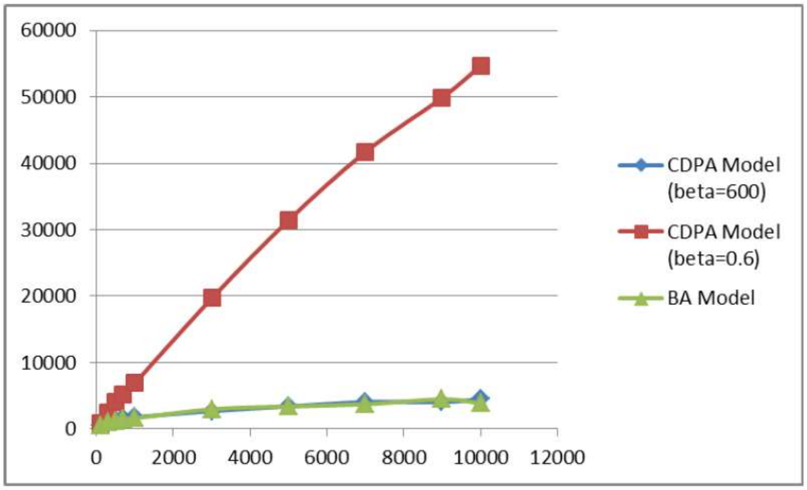


Figure 3.10 : Number of triangles is plotted for different size of networks generated by CDPAM with different values of β when $\theta = 0.5$. Horizontal-axis represents the number of nodes of the networks and the vertical-axis represents number of triangles in the networks.

if G is connected. More importantly, $\lambda_2(L)$ determines the robustness of a network i.e. larger the value of $\lambda_2(L)$, the more difficult to make the network disconnected by removal of nodes or edges [Fiedler, 1973]. In particular, if $\mu(G)$ and $\eta(G)$ denote the vertex and edge connectivity of a network G respectively, then $\lambda_2(L) \leq \mu(G) \leq \eta(G)$. In Fig. 3.11, we show that the complex networks generated by CDPAM setting $\beta \approx \theta$ have higher algebraic connectivity than that of networks produced by the BA model.

3.2.5 Spectral radius

Spectral radius of a network is the maximum of the absolute values of eigenvalues of the network. In [Jamakovic, 2008] it has been shown that the reciprocal of the spectral radius decides the threshold of virus propagation in the network. The smaller the spectral radius is, the larger the robustness of a network against the spread of viruses [Jamakovic, 2008]. In Fig. 3.12, we plot the spectral radius of networks generated by CDPAM and by BA model. As it appears, CDPAM produces networks with higher spectral radii compared to the spectral radii of the networks generated by BA model. Thus, CDPAM is capable to inherit large spectral radius like many real world networks including Dutch soccer team network [Jamakovic, 2008], Dutch roadmap network [Jamakovic *et al.*, 2006], Internet graph at the IP-level [Jamakovic and Van Mieghem, 2006] and the Autonomous System level [Mühlbauer *et al.*, 2006].

3.3 CONCLUSION

In the literature of social choice theory and management science it has been established that the choice of a person gets influenced by a given offered set and ultimately, the choice is determined by the local and global contexts of the items in the offered set. Inspired by this concept, we introduced a preferential attachment model for growing complex networks when the preference of a new node to get linked with old nodes in a network is determined by local and global properties of the old nodes. We call the model, the context dependent preferential attachment model (CDPAM). We proved that the complex networks generated by CDPAM have power law degree distribution and expected diameter depends logarithmically with the size of

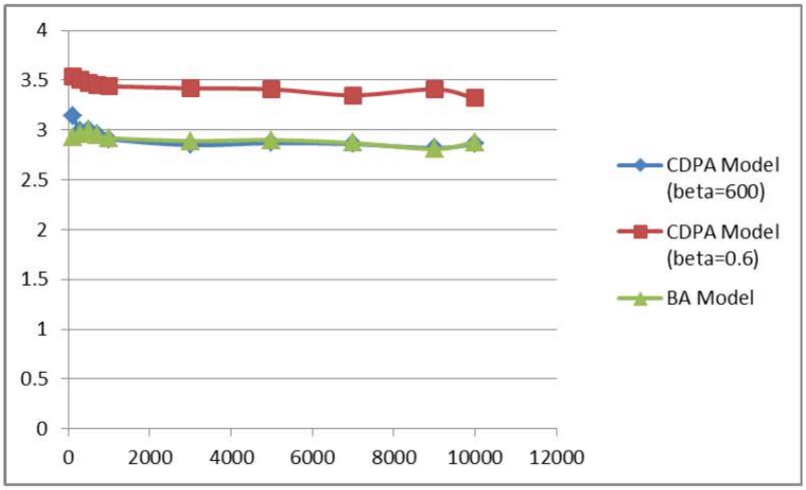


Figure 3.11 : Algebraic connectivity of different size of networks generated by CDPAM with different values of β when $\theta = 0.5$ and by BA model. Horizontal-axis represents the number of nodes of the networks and the vertical-axis represents algebraic connectivity of the networks.

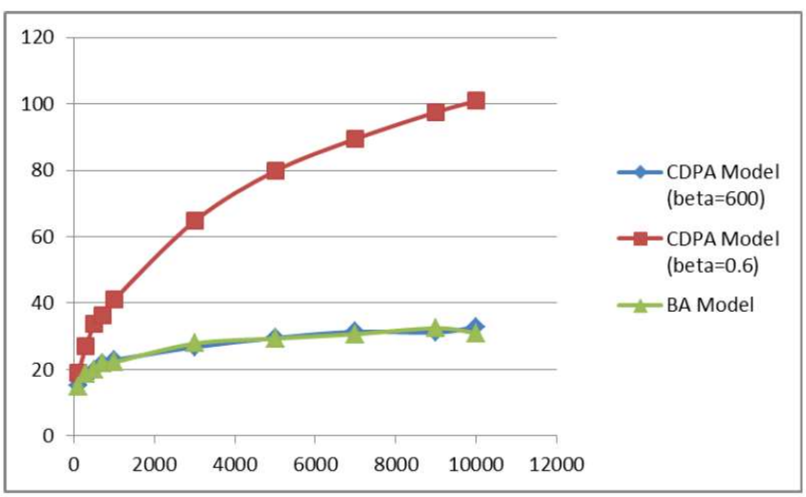


Figure 3.12 : Spectral radii for different size of networks generated by CDPAM with different values of β when $\theta = 0.5$ and by BA model. Horizontal-axis represents the number of nodes of the networks and the vertical-axis represents spectral radius of the networks.

new nodes added in the network during the growth process. In contrast to the general intuition that diameter grows with the addition of new nodes, we numerically showed that, in the CDPAM model, the expected diameter stabilizes when the new nodes get linked by giving alike importance (weight) to both local and global property of the old nodes.

In order to investigate how the complex networks generated by CDPAM and BA model are related, we calculated clustering coefficient, assortativity, number of triangles, algebraic connectivity and spectral radius for both the models. We compared these measures and concluded that BA model is a limiting case of CDPAM when new nodes tend to give large weight to the local property compared to the weight given to the global property during link formation.

…