A NOVEL USE OF STOCHASTIC APPROXIMATION ALGORITHMS FOR ESTIMATING DEGREE OF EACH NODE IN SOCIAL NETWORKS

Maziyar Hamdi *Vikram Krishnamurthy*

University of British Columbia,
Department of Electrical and Computer Engineering,
Vancouver, Canada, V6T 1Z4
{maziyarh, vikramk}@ece.ubc.ca

ABSTRACT

A duplication-deletion random graph is presented in this paper to model social networks which change over time. The paper analyzes the dynamics of this duplication-deletion random graph where at each time instant, one node can either join or leave the network. A degree distribution analysis is provided for this graph and an expression is derived to compute the power law component. Also a Markov-modulated random graph is analyzed where the the growth of the network evolves according to a slow Markov chain. An upper bound is derived for the mean square error between the estimated degree distribution and the asymptotic one. Using the fact that the duplication-deletion graph satisfies a power law, an upper bound is presented for the most significant singular value of the adjacency matrix of the graph.

***Index Terms*—** Complex networks, Markov modulated random graphs, power law, stochastic approximation.

1. INTRODUCTION

Social networks pervade the social and the economic lives of many people nowadays. They have a significant role in transmission of information through a network or even in spread of a disease or propagation of gossip in a community. Social networks are also important in several aspects of our daily life such as advertising goods and products, how we vote, which item we buy, etc. Because of the wide variety of application of social networks, finding a model which suits the social network the best and analyzing this model has attracted many attention recently.

Dynamic random graphs have been used widely to model social networks. In dynamic random graphs, the network evolves over time and a node can join the network at any time. Such dynamic models can be viewed as an infinite sequence of graphs where the random graph at each time may depend on all the earlier graphs (snapshots of the evolving graph at earlier times)[1]. In social networks, each node represents a member of this network and each edge depicts a relation between the incident nodes.

The evolution of the random graphs is investigated in several papers, such as [2, 3]. The model of Pastor-Satorras et al.[4] makes the basis for the model which is studied and generalized in this paper. In the Pastor-Satorras model, at each time step, a new node joins the network. In the literature, it has been shown that the degree distribution of such network satisfies a *power law*[5, 6]. In random graphs which satisfy the power law, the number of nodes with an specific degree depends on a parameter called power law component. In this paper, we generalize the model in [4, 1] to a scenario that each node can join or leave the network at any time (duplication-deletion model). The graph resulted from the duplication-deletion process is used to model social networks where the interaction between nodes evolves over time, e.g., the high school friendship (social) network whose growth is varying over time. To model such social networks, we assume that the evolution of the graph is changing according to a finite state Markov chain. This means that the probability of having new edges between nodes in the graph changes over the time. The proposed dynamical graph is very general and can cover more realistic social networks. A class of stochastic approximation (SA) algorithms is employed to track the cumulative distribution function (CDF) of degree of each node in the duplication-deletion random graph [7]. The evolution of duplication-deletion random graph in this case, depends on a slow Markov chain, therefore, the asymptotic behavior of such graphs is analyzed using a regime switching stochastic approximation algorithm. knowing the total number of nodes with specific degree (degree sequence) of the graph which models the social network, is important in analyzing the behavior of this network, for example it determines the existence of "giant component" from which, information transmission in a social networks can be investigated. The existence of giant component is also used in studying the spread of a disease through a human-related network, see [8, 9, 10].

Main Results: In this paper, a general duplication-deletion random graph is presented to model the social networks. This duplication-deletion process can be used to model online social networks in which nodes can join or leave the network at any time. We show that the graph resulted from duplication-deletion process, satisfies a power law. An equation for finding the power law components, β, for this model is presented in Sec.2.2. Singular values of the adjacency matrix of the resulting graph which have several applications in networks such as rank reduction and data mining, is also studied in this paper. This model is extended to a scenario that the evolution of the graph is a function of a Markov chain. Using the stochastic approximation algorithms, an equation is derived for the empirical measure of cumulative distribution function of degree of each node. Finally, it has been shown in this paper that the mean square error between estimated CDF and the expected one is bounded and is in the order of step size.

The remainder of the paper is organized as follows. In Sec.2.2, we provide a degree distribution analysis of the duplication-deletion random graph. An equation for the power law component and a discussion on the most significant singular value are also presented in this section. The Markov modulated random graph and the stochastic approximation algorithm to estimate the CDF are described in Sec.3. Numerical examples are given in Sec.4. Finally, Sec.5 concludes the paper and provides possible directions for future work.

2. DUPLICATION-DELETION RANDOM GRAPH

In this section, the duplication-deletion model is described in details and then a theorem is provided to prove that the resulting graph

from the duplication-deletion process satisfies a power law. An equation is also presented to compute the power law component for such duplication-deletion random graph. Proofs of theorems are omitted from this manuscript due to the lack of space. However, the sketches of the proofs are provided to give some intuition of different steps of the proofs.

2.1. Model description

Let $t = 0, 1, 2, \ldots$ denote discrete time. At each time, an arbitrary node joins the network by connecting to an specific node (parent node). After this step, all neighbors of the parent node (that is, all nodes connected to the parent node) connect to the new node with probability p. In this paper we assume that they are the only nodes who have chance to connect to the new node with a given probability. The model consists of two steps: *Duplication step* and *Deletion step*. In the duplication-deletion model there are three parameters: (i)p, probability of connection, (ii) q, probability of deletion, and (iii) G_0 an initial graph at time 0.

Given a graph G_t at time t, the dynamics evolve as follow:

Duplication step (occurs with probability 1):

- A node u from graph G_k is selected with uniform distribution.
- *Vertex-duplication*: New node v is generated. (A new vertex is added to the graph.)
- *Edge-duplication*:
  - Node v is connected to node u. (A new edge between u and v is added to the graph.)
  - With probability $p \in [0,1]$, node v is connected to each neighbor of node u.

Deletion Step (occurs with probability q):

- *Edge-deletion*: All the incident edges of a randomly chosen node (with uniform distribution), are deleted from the graph G_k.
- *Vertex-deletion*: The node which is chosen in the edge-deletion step is removed from the graph G_k as well.

The resulting graph is denoted by G_{t+1}. Therefore, the random graph evolves according to the duplication-deletion model as

$$G_{t+1} = \mathscr{G}(G_t, p, q), \tag{1}$$

where p and q are as defined above and the initial graph at time $t = 0$ is G_0.
The above dynamic random graph is used to model the social networks in where each node can join or leave at any time according to some probabilistic model.

2.2. Finding Power Law Component

Here, it is shown that the duplication-deletion random graph defined in Sec.2.1 satisfies a power law and an equation is presented for the power law component. We assume that the duplication step occurs first (before the deletion step) and the node generated in the duplication step cannot be eliminated in the deletion step immediately after its generation. This assumption makes the degree distribution analysis simpler and avoids the unrealistic situations. In spite of the deletion step, the duplication step occurs at each time with probability one in the duplication-deletion model. In each duplication step, the graph evolves with at least one more edge. If the probability of deletion, q, is small enough then the graph does not end up with a singleton. In the matrix form, the duplication step is adding a pair of row and column to the *adjacency matrix* of graph G. In the deletion step, with probability q, a random node j is uniformly chosen and the corresponding row and column are removed from the adjacency matrix of the graph.

Definition 2.1 *Let n_k denote the number of nodes of degree k in a random graph G_t in (1). Then G_t satisfies a power law distribution if n_k is proportional to $k^{-\beta}$ for a fixed $\beta > 1$: $\log n_k = \alpha - \beta \log k$, where α is a constant. β is called power law component.*

Let N_t denote the total number of nodes at time t. For simplicity we can assume that G_0 at time $t_0 = 0$ is an empty set so the graph at time $t = 1$ is a singleton. It is clear that if the probability of deletion, q, is zero then at time t, $N_t = t$; because at each step one node is added to the graph and no vertex is deleted. Let $f(t,i)$ denote the number of vertices with degree i at time t. Theorem2.1 gives an expression for the expected value of $f(t,i)$ and finds the power law component for the duplication-deletion random graph in terms of p and q.

Theorem 2.1 *With probability approaching 1, the duplication-deletion random network G_t, $t = 1, 2 \ldots$ defined in (1) satisfy a power law as $t \to \infty$. The power law component, β, can be computed from following equation.*

$$1 + p - \beta p - p^{\beta - 1} = q - \beta q, \tag{2}$$

where p and q are the probabilities defined in duplication and deletion steps.

Sketch of Proof: Considering all the events that result in a node with degree $i+1$, at time $t+1$, a recurrence formula is derived for conditional expectation of $f(t+1, i+1)$. Solving this recursive equation completes the proof.

In the duplication-deletion random graph considered above, if the maximum degree of a node is bounded, then the maximum degree of a random graph with power law component, β, is at most $e^{\frac{\alpha}{\beta}}$ where α is defined in Definition2.1, see [1].

The singular values of the adjacency matrix of a random graph and specifically the most largest ones have many applications in dealing with large graphs e.g. rank reduction, graph matching and link prediction in random graphs. Also singular value decomposition (SVD) method is widely used in information retrieval and data mining[11]. In this section, we are focusing on the singular values of the adjacency matric of the duplication-deletion model. Because the adjacency matrix of non-directional graph is symmetric, the singular values of the adjacency matrix s equal to the eigenvalues of that matrix[12, 13].

It has been shown in the literature that the eigenvalues of the adjacency matrix of a power law graph satisfy the characteristics described in Theorem2.2.[1] This theorem provides an upper bound for the largest eigenvalue of such matrices and also stated that the most significant eigenvalues of a power law random graph also has power law distribution under additional conditions on power law component and the maximum degree[14].

Theorem 2.2 ([14]) *In a random graph that satisfies power law with exponent β, if $\beta > 2.5$ then the largest eigenvalue of the adjacency matrix is bounded by $(1 + O(1))\sqrt{m}$ and the k largest eigenvalues also have power law distribution with exponent $2\beta - 1$, for $k < n\left(\frac{d}{m \log n}\right)^{\beta - 1}$, where n is the number of vertices and m and d are the maximum and the average degree, respectively.*

[1] The complete proof can be found in [14]

3. MARKOV MODULATED RANDOM GRAPH

This section generalizes the model and analysis of Sec.2.2. It considers the case where the probability of connection p evolves according to a finite state Markov chain, θ_t. As shown in Sec.2.2, the power law component depends on p. Therefore, as p evolves over time, the power law component varies with time. So in the Markov modulated duplication-deletion random graph, the power law component depends on θ_t, $\beta(\theta_t)$. Using a stochastic approximation algorithm for the cumulative distribution function, the aim is to estimate the cumulative distribution function of each node's degree. As defined in Sec.2.2, $f(t,i)$ is the number of nodes with degree i so $\sum_{k=1}^{\infty} f(t,i) = N_t$. Therefore, $\frac{f(t,i)}{N_t}$ can be interpreted as a probability mass function of random process, x_t, which denotes the degree of an specific node at time t.

Let $\{\theta_t\}$ be a discrete-time slow Markov chain with finite state space

$$\mathscr{M} = \{\bar{\theta}_1, ..., \bar{\theta}_{m_0}\}, \tag{3}$$

and transition probability matrix

$$A^{\varepsilon} = I + \varepsilon Q. \tag{4}$$

Here ε is a small parameter and I is an $m_0 \times m_0$ identity matrix, and Q is an irreducible generator of a continues-time Markov chain. Let q_{ij} denote the elements of the generator matrix Q such that

- (A) $q_{ij} \geq 0$ if $i \neq j$ and $\forall i$, $\sum_{j=1}^{m_0} q_{ij} = 0$.

The assumption of irreducibility implies that there exists a unique stationary distribution for this Markov chain, $\pi \in \mathbb{R}^{m_0 \times 1}$ such that

$$\pi' = \pi' A^{\varepsilon}. \tag{5}$$

The total number of nodes with degree i at time t in the Markov modulated random graph depends on the state of the Markov chain θ_t. Let $\bar{f}(t,i,\theta_t)$ denote the total number of nodes with degree i at time t. The new parameter $g_t(n,\theta_t)$ is defined as

$$g_t(n,\theta_t) = \frac{1}{N_t} \sum_{k=1}^{n} \bar{f}(t,k,\theta_t), \tag{6}$$

where can be interpreted as the cumulative distribution function of degree of nodes at time t in the Markov modulated random graph. In a Markov-modulated duplication-deletion random graph, the expected CDF is varying over time as the state of Markov chain changes. When t is sufficiently large, the expected value of $g_t(n,\theta_t)$ can be written as

$$\begin{aligned}
\mathbf{E}\{g_t(n,\theta_t)\} &= \mathbf{E}\left\{\sum_{i=1}^{m_0} I(\theta_t = \bar{\theta}_i).g_t(n,\bar{\theta}_i)\right\} \\
&= \sum_{i=1}^{m_0} I(\theta_t = \bar{\theta}_i).\mathbf{E}_{\theta}\{g_t(n,\bar{\theta}_i)\} \\
&= \sum_{i=1}^{m_0} I(\theta_t = \bar{\theta}_i).\sum_{k=1}^{n} Ck^{-\beta(\bar{\theta}_i)} \\
&= C\sum_{i=1}^{m_0}\sum_{k=1}^{n} \pi(i)k^{-\beta(\bar{\theta}_i)},
\end{aligned} \tag{7}$$

where π is defined in (5). Here, it is shown that if the Markov chain is slow enough, the stochastic approximation algorithm with constant step size is still able to estimate the cumulative distribution function and the estimated cumulated distribution function follows the expected one precisely.

As $t \to \infty$ in (1), the support of the degree distribution becomes unbounded in general. But in a power law random graph (recall from Sec.2.2), the maximum degree does not depend on time. Let M denote the maximum degree of the power law random graph. We assume that g_t is a $1 \times M$ vector. The i-th element of $g(t)$ can be found from (6).

Let $X_t \in \mathbb{R}^M$ denote the observed degree sequence of the resulted graph from the duplication-deletion process at time t. These local observations are used to estimate the empirical cumulative distribution function. The empirical measure of cumulative distribution function is defined as follows,

$$\hat{g}_t(n) = \frac{1}{t}\sum_{k=0}^{t-1} I_{\{X_k(n) \leq n\}}. \tag{8}$$

$\hat{g}(t,n)$ can be written recursively as follows,

$$\hat{g}_{t+1}(n) = \hat{g}_t(n) - \frac{1}{t+1}\hat{g}_t(n) + \frac{1}{t+1}\left(I_{\{X_{t+1}(n) \leq n\}} - \hat{g}_t(n)\right). \tag{9}$$

If t is sufficiently large, the following stochastic approximation algorithm with constant step size, μ (where μ denotes a small positive constant), is used to estimate the empirical cumulative distribution function,

$$\hat{g}_{t+1}(n) = \hat{g}_t(n) + \mu\left(I_{\{X_{t+1}(n) \leq n\}} - \hat{g}_t(n)\right). \tag{10}$$

Here, we study the asymptotic behavior of the expected degree distribution. We show that the difference between the expected CDF and the estimated one is bounded and this bound depends on μ and ε. This means that the empirical CDF follows the expected CDF properly and the error between these two is bounded. Let $\tilde{g}_t(n) = \hat{g}_t(n) - \mathbf{E}\{g_t(n,\theta_t)\}$. Theorem3.1 shows that the difference between sample path and the expected cumulative distribution function is bounded. It also finds the order of this difference in terms of μ and ε.

Theorem 3.1 [2] *Suppose that $\varepsilon^2 = o(\mu)$[3], then for sufficiently large t,*

$$\mathbf{E}|\tilde{g}_t|^2 = O\left(\mu + \varepsilon + \frac{\varepsilon^2}{\mu}\right). \tag{11}$$

Sketch of Proof: We first define a Lyapunov function $V(\tilde{g}(t)) = \frac{1}{2}\tilde{g}(t)\tilde{g}'(t)$. Then, a recursive expression is written for the growth of the difference between sample path and the expected CDF, $V(\tilde{g}_{t+1}) - V(\tilde{g}_t)$. Back-ward iterating and interpolation conclude the proof of the theorem.

The above theorem implies that the mean square error between the expected cumulative distribution function and the empirical one is bounded. Therefore for small μ, the empirical distribution is an accurate estimate of the expected CDF. The expected CDF can be used, for example, in finding the probability of having a giant component is social networks (which has many implications in social networks as described in Sec.1).

4. NUMERICAL EXAMPLES

In this section, numerical examples are given to illustrate the results from Sec.3. We start with implementing duplication and deletion steps in the scenario that the probability of connection is not changing over time. The resulting duplication-deletion random graph is investigated in terms of degree distribution to show that it satisfies

[2] Proofs for theorems are omitted due to the lack of space.

[3] Note that in many cases it is assumed that $\varepsilon = O(\mu)$ and therefore, $\varepsilon^2 = o(\mu)$ is a consequence.

power law. Theorem2.1implies that the degree sequence of the resulted graph satisfies power law with exponent computed using (2). Let β^* denote the solution of (2). $\beta^* = 1$ always satisfies (2). Power law component, $\beta = \max\{1, \beta^*\}$.

Fig.1 Shows the non-trivial solution of (2) versus p for different values of probability of deletion, q. As can be seen in this figure, if the probability of deletion is relatively high then the power law component is very large and this means that a majority of the nodes in this graph has smaller degree and few nodes has larger degree.

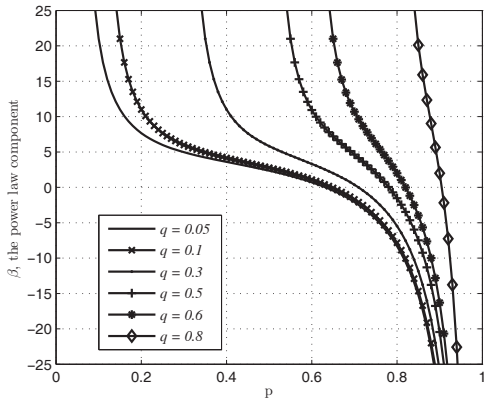

Fig. 1. The non-trivial solution of the equation 2 for different values of p and q

Fig.2 shows the number of nodes with specific degree on a logarithmic scale for the both horizontal and vertical axes for the graph resulted from duplication-deletion process with probabilities of connection and deletion as follows, $p = 0.48$ and $q = 0.1$. It can be inferred from the linearity in Fig.2 (excluding the nodes with very small degree), that the resulted graph from duplication-deletion process satisfies a power law. Also, the slope of the linear part in Fig.2 suggests a value for the power law component which is close to that obtained from Fig.1 ($\beta^* = 3.04$). As can be seen in the Fig.2, the power law is a better approximation for the middle points compared to both ends.

The numerical example for the Markov-modulated random graph which corroborate the result of Theorem3.1 is omitted due to the lack of space.

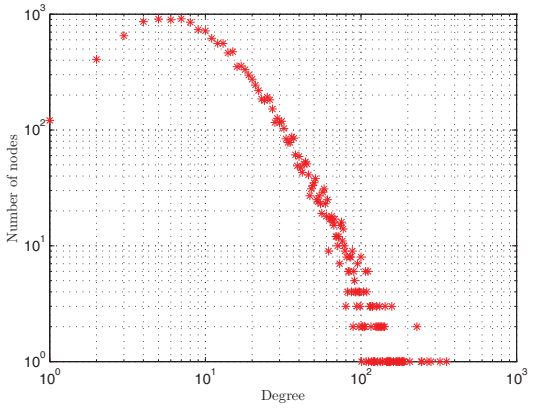

Fig. 2. The degree distribution of the duplication-deletion random graph in log-log scale.

5. CONCLUSION AND FUTURE WORKS

This paper analyzed the dynamics of a duplication-deletion graph where at each time instant, one node can either join or leave the graph (an extension to the duplication model of [15, 4]). The power law component for such graph was computed using the result of Theorem2.1. Also the Markov modulated random graph was proposed to model the social networks whose evolution changes over time. Using the stochastic approximation algorithms, the cumulative distributions function of degree of each node is estimated. Finally, an upper bound was derived for the distance between the empirical and the expected CDF. Characterizing the error between the expected CDF and the estimated one can be an extension of this work. Investigating the spread of a message through a social network using the expected degree distribution presented in this paper, can also be a topic for future work.

6. REFERENCES

[1] F. Chung and L. Lu, *Complex Graphs and Networks*. Conference Board of the Mathematical Sciences, National Science Foundation (U.S.), 2006.

[2] P. Erdos and A. Renyi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[3] E. Lieberman, C. Hauert, and M. A. Nowak, "Evolutionary dynamics on graphs," *Nature*, vol. 433, no. 7023, pp. 312–316, Jan. 2005.

[4] R. Pastor-Satorras, E. Smith, and R. V. Sol, "Evolving protein interaction networks through gene duplication," *Journal of Theoretical Biology*, vol. 222, no. 2, pp. 199 – 210, 2003.

[5] H. Jeong, S. P. Mason, A. L. Barabsi, and Z. N. Oltvai, "Lethality and centrality in protein networks." *Nature*, vol. 411, no. 6833, pp. 41 – 42.

[6] A. Wagner, "The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes," vol. 18, no. 7, pp. 1283–1292, 2001.

[7] H. Kushner and G. Yin, *Stochastic approximation algorithms and applications*, ser. Applications of mathematics. Springer-Verlag, 1997.

[8] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. Suppl 1, pp. 2566–2572, 2002.

[9] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, p. 208701, Oct 2002.

[10] S. Eubank, H. Guclu, V. S. Anil Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, pp. 180–184, May 2004.

[11] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, vol. 37, no. 4, pp. pp. 573–595.

[12] J. N. Franklin, *Matrix Theory*. Dover Publications, 1993.

[13] M. Hazewinkel, *Enclopedia of Mathematics*. Springer, 2001.

[14] M. Mihail and C. Papadimitriou, "On the eigenvalue power law," in *Randomization and Approximation Techniques in Computer Science*.

[15] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas, "Duplication models for biological networks," *Journal of Compuational Biology*, vol. 10, pp. 677–687, 2003.