EuroComb 2005 DMTCS proc. **AE**, 2005, 239–244

Degree-correlation of a Scale-free Random Graph Process

Zoran Nikoloski[1†] and Narsingh Deo [1] and Ludek Kucera [2]

[1]*School of Computer Science, University of Central Florida, Orlando, FL 32816, USA*
[2]*Department of Applied Mathematics, Faculty of Physics and Mathematics, Charles University, Prague, Czech Republic*

Barabási and Albert [1] suggested modeling scale-free networks by the following random graph process: one node is added at a time and is connected to an earlier node chosen with probability proportional to its degree. A recent empirical study of Newman [5] demonstrates existence of degree-correlation between degrees of adjacent nodes in real-world networks. Here we define the *degree correlation*—correlation of the degrees in a pair of adjacent nodes—for a random graph process. We determine asymptotically the joint probability distribution for node-degrees, d and d', of adjacent nodes for every $0 \leq d \leq d' \leq n^{1/5}$, and use this result to show that the model of Barabási and Albert does not generate degree-correlation. Our theorem confirms the result in (KR01), obtained by using the mean-field heuristic approach.

please also repeat in the submission form

Keywords: degree-correlation, scale-free degree distribution, linearized chord diagrams

1 Introduction

The Internet and the World Wide Web (WWW), at first envisioned as technological networks for dissemination of scientific information, now serve various commercial purposes. The appearance of certain global characteristics in these real-world networks can be determined by using random graph processes; for instance, the WWW can be modeled by a graph where a node represents a web-page, while the edges are the hyperlinks connecting the web-pages. Empirical studies of the Internet and the WWW have shown statistical similarities between these and other networks in terms of: the scale-free degree distribution, clustering coefficient, and the average distance (Mit04; WS98). Developing random-graph models that closely match the characteristics of real-world networks is the first step to designing (or transforming) a network in a way that a given purpose ($e.g.$, reliable searching) could be reached in an efficient way.

Recent empirical studies of technological and social networks [6] demonstrated correlation among the degrees of adjacent nodes—*i.e.*, *degree-correlation*. Here, we define the degree-correlation for scale-free random graph processes. Our approach is similar to that employed in (BO04; BR04) and confirms that node-degrees in the Barabási-Albert model (BA99) *are not* correlated.

[†]Partially supported by the EU DELIS Project - Charles University site

2 The model and the Pearson correlation coefficient

Bollobás and Riordan (BR04) gave a mathematically precise definition of the process introduced by Barabási and Albert (BA99). Consider a fixed sequence of nodes $v_1, v_2, \ldots$. The process $(G_1^t)_{t\geq 0}$ is inductively defined, as follows: G_1^1 is composed of one node and one loop. Given G_1^{t-1}, G_1^t is obtained by adding a node v_t together with a single edge between v_t and v_i, where i is randomly chosen with probability:

$$P(i=k) = \begin{cases} \frac{d_{G_1^{t-1}}(v_k)}{2t-1}, & 0 \leq k \leq t-1 \\ \frac{1}{2t-1}, & k = t \end{cases} .$$

If the number of added edges, m, from v_t, is greater than one, the process $(G_m^t)_{t\geq 0}$ is obtained by running $(G_1^t)_{t\geq 0}$ on the sequence $v'_1, v'_2, \ldots$; that is, form a graph G_m^t from the graph G_1^{mt} by identifying the nodes $v'_1, v'_2, \ldots v'_m$ to form v_1, identifying $v'_{m+1}, v'_{m+2}, \ldots v'_{2m}$ to form v_2, and so forth.

This definition allows the dynamic graph process to be analyzed via its static description—*linearized chord diagram* (*LCD*) (Sto99): The linearized chord diagrams (LCD), with n chords, consist of $2n$ distinct points on the x-axis paired off by semi-circular chords, each chord having one left and one right endpoint. A graph can be obtained from an LCD as follows: starting from the left, identify all endpoints up to and including the first right endpoint reached from node 1. The rest of the nodes are obtained by repeating this process. Finally, the chords from the LCD represent edges in the obtained graph.

The Pearson correlation coefficient, r, is a real number, in the range [-1, 1], that expresses the quality of the least square fitting to a given set of data points (x_i, y_i), $1 \leq i \leq n$. There are two evident problems: (1) how to choose which of the degrees in a pair of adjacent nodes to represent x_i and y_i, and (2) the correlation coefficient should asymptotically hold for any graph generated by the random graph process. Here, it is more convenient to use the correlation coefficient, r, for two random variables X and Y, written as $r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where $cov(X,Y) = E[(X,Y)] - E[X]E[Y]$, and (X,Y) represents the joint probability distribution of the random variable X and Y.

Given a random graph G_m^n generated by the random graph process $(G_m^t)_{t\geq 0}$, consider the two-stage experiment: (1) choose an edge $e = (u,v)$ from G_m^n independently at random, (2) choose one node, say u, incident with e. Let $d(u)$ be the value of X, and $d(v)$ be the value of Y. The probability distribution of the random variable X can be easily derived. Let the number of d-degree nodes be N_d. Since each edge results in two possibilities for successful events, one can obtain the following: $P(X=d) = \frac{dP(Z=d)}{\sum_d dP(Z=d)}$, where $P(Z=d)$ is the probability distribution of the r. v. representing the degree of a node chosen uniformly at random. Clearly,X and Y have the same probability distribution. For convenience, we use the abbreviated notation: $P(X=d) = q_d$, $P(Z=d) = p_d$, and $P(X=d, Y=d') = e_{dd'}$. The Pearson correlation coefficient can be calculated as:

$$r = \frac{\sum\limits_{d,d'} dd'\left(e_{dd'} - q_d q_{d'}\right)}{\sum\limits_d d^2 q_d - \left[\sum\limits_d d q_d\right]^2}.$$

To make use of this formulation, we need to derive an expression for $e_{dd'}$.

3 Joint Probability Distribution

We derive an expression for the joint-probability distribution for degrees of adjacent nodes, writing $\#_m^n(d,d')$ for the number of adjacent pairs of nodes with in-degrees d and d', $i.e.$ with total degree of $(m+d)$ and $(m+d')$. Theorem 1, below, shows the the result in [4] is correct.

Theorem 1. *Let* $m = 1$ *and* $(G_1^n)_{n\geq 0}$ *be the random graph process defined in Section 2. Let*

$$\alpha_{d,d'} = \frac{\frac{4\left(d'-1\right)}{d(d+1)(d+d')(d+d'+1)(d+d'+2)}+}{\frac{12\left(d'-1\right)}{d(d+d'-1)(d+d')(d+d'+1)(d+d'+2)}},$$

and let $\varepsilon > 0$ *be fixed. Then with probability tending to 1 as* $n\to\infty$ *we have*

$$(1-\varepsilon)\,\alpha_{d,d'} \leq \frac{\#_1^n\,(d,\,d')}{n^2} \leq (1+\varepsilon)\,\alpha_{d,d'}$$

for every $0\leq d\leq d'\leq n^{1/5}$.

Proof: It turns out that we only need to calculate the expectation $\#_m^n(d,d')$; the concentration result is then given by applying the Azuma-Hoeffding inequality. The strategy of the proof is as follows: It is enough to consider the case when $m=1$, the result for general m follows, as mentioned. First, we derive explicitly the joint distribution of D_k and $D_{k'}$, where D_k (resp. $D_{k'}$) is the sum of the first k (resp. k') degrees, assuming $k' > k$. Bollobás and Riordan (BR04) already proved that D_k is concentrated about a certain value. We combine these results to obtain approximately the joint probability ($d_{G_1^n}(v_{k+1}) = d+1,\ d_{G_1^n}(v_{k'+1}) = d'+1$). Summing over k and k' gives the desired result.

Consider first the event $\{D_k - 2k = s\}$, where $0\leq s\leq n-k$. This is the event that the last $n-k$ nodes of G_1^n send exactly s edges to the first k nodes. This corresponds to a LCD in which the kth right endpoint is $2k+s2k+s$. We shall split the this LCD into left partial LCD L, induced on $\{1,\ldots,2k+s\}$, and a right partial LCD R, induced on $\{2k+s+1,\ldots,2n\}$. Similarly, we arrive at the partial LCDs L' and R', generated by the event $\{D_{k'}-2k'=s'\}$, where $0\leq s'\leq n-k'$. Suppose that the left partial LCDs L and L' share j common left unpaired endpoints, where $0\leq j\leq \min(s,s')$. Consider the event $\{D_k - k = s, D_{k'}-k' = s'|j\}$, the corresponding left partial LCD has exactly

$$\Psi\frac{(2n-2k'-s')!}{(2n-2k'-2s')!}\frac{(2k'-2k-s+j)!}{(2k'-2k-2s+2j)!}\,(2n-2k-2s-s'+j-3)!!$$

extensions to a full n-pairing. The term Ψ denotes a rather unilluminating expression that simplifies to $ss'-j$.

This extension of the left partial LCD corresponds to a graph with $d_{k+1} = d+1$ and $d_{k'+1} = d'+1$ if and only if $2k+s+d+1$ and $2k'+s'+d'+1$ are right endpoints, and each of the $2k+s+1,\ldots,2k+s+d, 2k'+s'+1,\ldots,2k'+s'+d'$ is a left endpoint. Note that the element paired with $2k+s+d+1$ must be either one of the s unpaired elements in L or one of the $2k+s+1,\ldots,2k+s+d$, and that $s+d-1$ pairs start before $2k+s+d+1$ and end after this point. In order for v_{k+1} and $v_{k'+1}$ to be adjacent, it is easy to conclude that $2k'+s'+d'+1$ must only be paired with one of the unpaired $2k+s+1,\ldots,2k+s+d$, and that $s'+d'-1$ pairs start before $2k'+s'+d'+1$and end after this point. Since we also have to consider the number j of overlapping unpaired endpoints in the left partial LCDs L and L', we arrive at three cases: (1) $2k+s+d+1$ chooses among j overlapping left endpoints, $2k'+s'+d'+1$ chooses among d unpaired left endpoints immediately preceding $2k+s+d+1$, (2) $2k+s+d+1$ chooses among $s-j$ non-overlapping left endpoints, $2k'+s'+d'+1$ makes the same choice as in the previous case, and (3) Each of $2k+s+d+1$ and $2k'+s'+d'+1$ chooses one left endpoint from $2k+s+1,\ldots,2k+s+d$. Such left partial LCD has exactly

$$\Upsilon \tfrac{(2n-2k'-s'-d'-1)!}{(2n-2k'-2s'-2d'-1)!} \tfrac{(2k'-2k-s-d+j-1)!}{(2k'-2k-2s-d+2j-1)!} \cdot \\ \cdot \tfrac{(2n-2k-2s-s'-d-d'+j-1)!}{(2n-2k-2s-s'-2d-d'+j-1)!} (2n-2k-2s-s'-2d-d'+j-4)!!$$

extensions to a full n-pairing. The term Υ denotes a rather unilluminating expression that simplifies to $d(d+s-1)$. Let $M = \lfloor n^{4/5}/\log n \rfloor$, let $k = k(n)$ (resp. $k' = k'(n)$) be any function satisfying $M \leq k(n) \leq n-M$, and let $d = d(n)$ and $d' = d'(n)$ be any two functions satisfying $0 \leq d'(n) \leq d(n) \leq n^{1/5}$. One may obtain:

$$P(d_{k+1} = d, d_{k'+1} = d' | D_k - k = s, D_{k'} - k' = s', j) = \\ (1+o(1)) \left[\tfrac{2(\sqrt{n}-\sqrt{k})^2}{2(n-\sqrt{kn})} \right]^{2d'+1} \left[\tfrac{2(k'+k-2\sqrt{kn}+j)}{2k'-2\sqrt{kn}+j} \right]^{d+1} = \\ (1+o(1)) \left(1-\sqrt{\tfrac{k}{n}}\right)^{2d'+1} \left(1-\sqrt{\tfrac{k}{n}}+1-\sqrt{\tfrac{k'}{n}}\right)^{d+1}$$

Thus, we arrive at:

$$E[\#_1^n(d,\, d')] \sim \sum_{k'=M}^{n-M} \sum_{k=M}^{n-M} \left(1-\sqrt{\tfrac{k}{n}}\right)^{2d'+1} \left(1-\sqrt{\tfrac{k}{n}}+1-\sqrt{\tfrac{k'}{n}}\right)^{d+1} \\ = n^2 \int_0^1 \left[\int_0^1 (1-\sqrt{\kappa})^{2d'+1} \left(1-\sqrt{\kappa}+1-\sqrt{\kappa'}\right)^{d+1} d\kappa \right] d\kappa'$$

where $\kappa = k/n$ and $\kappa' = k'/n$. The inner integral yields

$$\int_0^1 (1-\sqrt{\kappa})^{2d'+1} \left(1-\sqrt{\kappa}+1-\sqrt{\kappa'}\right)^{d+1} d\kappa = \\ \tfrac{(1-\sqrt{\kappa'})^d}{(3+5d'+2d'^2)} \left(\begin{array}{l} \left(1-\sqrt{\kappa'}\right)(3+2d')\,{}_2F_1\left(2+2d',\ -d,\ 3+2d',\ -\tfrac{1}{1-\sqrt{\kappa'}}\right) + \\ +(2+2d')\,{}_2F_1\left(3+2d',\ -d,\ 4+2d',\ -\tfrac{1}{1-\sqrt{\kappa'}}\right) \end{array} \right)$$

which integrated over κ' gives

$$E[\#_1^n(d,\ d')]/n^2 \sim \\ \tfrac{(6+4d')\Gamma(-2-d)\Gamma(3+2d')(1+d)\,{}_2F_1(-2-d,2+2d',3+2d',-1)}{(3+10d')\Gamma(-d)} \\ -\tfrac{(12+8d')\Gamma(-2-d)\Gamma(3+2d')(2+d)(1+d')\,{}_2F_1(-1-d,3+2d',4+2d',-1)}{(3+10d')\Gamma(-d)} \\ -\tfrac{(4+4d')\Gamma(-1-d)\Gamma(4+2d')\,{}_2F_1(3+2d',-1-d,4+2d',-1)}{(3+2d')\Gamma(-d)} .$$

By using the Kummer's formula, the theorem follows.

References

[BA99] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[BO04] P. G. Buckley and D. Osthus. Popularity based random graph model leading to a scale-free degree sequence. *Discrete Mathematics*, 282:53–68, 2004.

[BR04] B. Bollobás and O. M. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1), 2004.

[KR01] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 066123, 2001.

[Mit04] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[New02] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 208701, 2002.

[Sto99] A. Stoimenow. On enumeration of chord diagrams and asymptotics of vassiliev invariants. *Ph.D. Thesis*, 1999.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.