Some typical properties of the Spatial Preferred Attachment model

Colin Cooper, Alan Frieze, and Pawel Pralat

[1] Department of Computer Science, King's College, University of London, London WC2R 2LS, UK
[2] Department of Mathematical Sciences, Carnegie Mellon University, 5000 Forbes Av., 15213, Pittsburgh, PA, U.S.A
[3] Department of Mathematics, Ryerson University, Toronto, ON, Canada, M5B 2K3

Abstract. We investigate a stochastic model for complex networks, based on a spatial embedding of the nodes, called the Spatial Preferred Attachment (SPA) model. In the SPA model, nodes have spheres of influence of varying size, and new nodes may only link to a node if they fall within its influence region. The spatial embedding of the nodes models the background knowledge or identity of the node, which influences its link environment. In this paper, we focus on the (directed) diameter, small separators, and the (weak) giant component of the model.

1. Introduction

Discrete random graph processes exhibiting power law properties have been studied by many authors and in many contexts. The study of such processes dates back at least, to Yule [30] in 1924. Recent interest in preferential attachment models follows from the work of Barabási and Albert [5] who observed a power law degree sequence for a subgraph of the World Wide Web, and of Faloutsos, Faloutsos and Faloutsos [16] who observed power law behaviour for the internet graph. Many models of such process exist. For details see, for example, the surveys [7, 29] and the monographs [9, 13].

In networked information spaces, vertices are not only defined by their link environment, but also by the information entity they represent. More recently, attempts have been made to model this alternative view of the vertices through *spatial models*. In a spatial model, vertices are embedded in a metric space, and link formation is influenced by the metric distance between vertices. The metric space is meant to be like a feature space, so that the coordinates of a vertex in this space represent the information associated with the vertex. For example, in text mining, documents are commonly represented as vectors in a word space. The metric is chosen so that metric distance represents similarity, i.e. vertices whose information entities are closely related will be at a short distance from each other in the metric space. A number of spatial models have been proposed up to date [10, 11, 17–19, 26]. We direct the reader to the recent survey for more details [20].

We focus on the Spatial Preferred Attachment (SPA) model, proposed in [3, 4]. The SPA model generates directed graphs according to the following principle. Vertices are points in a given metric space. Each vertex v has a *sphere of influence*. The volume of the sphere of influence of a vertex is a function of its in-degree. A new vertex u can only link to an existing vertex v if u falls inside the sphere of influence of v. In the latter case, u links to v with probability p. The SPA model incorporates the principle of preferential attachment, since vertices with a higher in-degree will have a larger sphere of influence. The SPA model gives a power law in-degree distribution, with exponent in $[2, \infty)$ depending on the parameters, and with concentration for a wide range of in-degree values [3, 4]. In [22, 21] it was shown, through theoretical analysis and simulation, that for graphs formed according to the SPA model it is possible to infer the metric distance between vertices from the link structure of the graph.

In this paper, we investigate the (directed) diameter, small separators, and the (weak) giant component of the model. This is an extended version of the paper presented at the 9th Workshop on Algorithms and Models for the Web Graph (WAW 2012) [14].

2. The SPA model

We start by giving a precise description of the SPA model, presenting some known properties, and deriving some facts about the model, which we will need to prove our results. In [3] (see also [4] for a proceeding version of this paper), the model is defined for a variety of metric spaces S. In this paper, we let S be the unit hypercube in $\mathbb{R}^m$, equipped with the torus metric derived from any of the L_p norms. This means that for any two points x and y in S,

$$d(x, y) = \min\left\{||x - y + u||_p \,:\, u \in \{-1, 0, 1\}^m\right\}.$$

The torus metric thus "wraps around" the boundaries of the unit square; this metric was chosen to eliminate boundary effects. m-th power of the radius in m dimensions, so the volume of a ball of radius r in m-dimensional space with the given metric equals $c_m r^m$. For example, for the Euclidean metric, $c_2 = \pi$, and for the product metric derived from L_∞, $c_m = 2^m$.

The parameters of the model consist of the *link probability* $p \in [0, 1]$, and two positive constants A_1 and A_2, which, in order to avoid the resulting graph becoming too dense, must be chosen so that $pA_1 < 1$. The original model as presented in [3] has a third parameter, A_3, which is assumed to be zero here. This causes no loss of generality, since all asymptotic results presented here are unaffected by A_3.

The SPA model generates stochastic sequences of graphs $(G_t : t \geq 0)$, where $G_t = (V_t, E_t)$, and $V_t \subseteq S$. Let $\deg^-(v, t)$ be the in-degree of vertex v in G_t, and $\deg^+(v, t)$ its out-degree. We define the *sphere of influence* $S(v, t)$ of vertex v at time $t \geq 1$ to be the ball centered at v with volume $|S(v, t)|$ defined as follows:

$$|S(v, t)| = \frac{A_1 \deg^-(v, t) + A_2}{t}, \tag{2.1}$$

or $S(v,t) = S$ and $|S(v,t)| = 1$ if the right-hand-side of (2.1) is greater than 1.

The process begins at $t = 0$, with G_0 being the null graph. Time-step t, $t \geq 1$, is defined to be the transition between G_{t-1} and G_t. At the beginning of each time-step t, a new vertex v_t is chosen *uniformly at random* from S, and added to V_{t-1} to create V_t. Next, independently, for each vertex $u \in V_{t-1}$ such that $v_t \in S(u, t-1)$, a directed link (v_t, u) is created with probability p. Thus, the probability that a link (v_t, u) is added in time-step t equals $p\,|S(u, t-1)|$.

We say that an event holds asymptotically almost surely (a.a.s.) if the probability that it holds tends to one as t goes to infinity. It was shown in [3] that a.a.s. the SPA model produces graphs with a power law degree distribution, with exponent $1 + 1/(pA_1)$. Moreover, a precise expression for the probability distribution of the in-degree of the individual vertex v_i born at time i was given. In [21] (see also [22]) the relationship between the link structure of graphs produced by the model and the relative positions of the vertices in the metric space was analyzed. See Figure 1 for a drawing of a simulation of the SPA model.

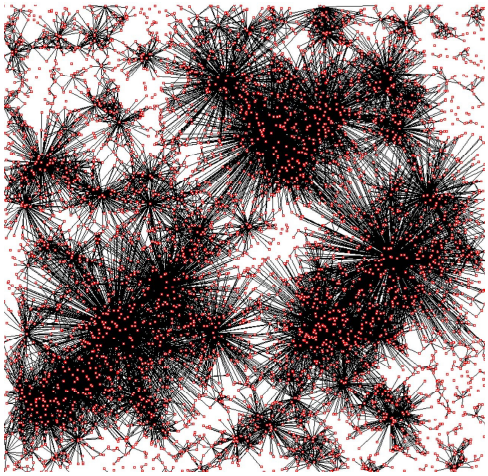


Fig. 1. A simulation on the unit square with $t = 5,000$ and $p = A_1 = A_2 = 1$.

Now, let us discuss a few simple new facts about the model. Knowing the expected in-degree of a node, given its age, will help us to analyze geometric properties of the SPA Model. Let us note that the result for $i \gg 1$ was proved in [21] (see (2.2)); we extend it here to all $i \geq 1$ (see (2.3)). As before, let v_i be the node added at time i.

Theorem 1. *Suppose that $i = i(t) \gg 1$ as $t \to \infty$. Then,*

$$\mathbb{E}(\deg^-(v_i,t)) = (1+o(1))\frac{A_2}{A_1}\left(\frac{t}{i}\right)^{pA_1} - \frac{A_2}{A_1}, \tag{2.2}$$
$$\mathbb{E}(|S(v_i,t)|) = (1+o(1))A_2 t^{pA_1-1} i^{-pA_1}.$$

Moreover, for all $i \geq 1$,

$$\mathbb{E}(\deg^-(v_i,t)) \leq \frac{eA_2}{A_1}\left(\frac{t}{i}\right)^{pA_1} - \frac{A_2}{A_1}. \tag{2.3}$$
$$\mathbb{E}(|S(v_i,t)|) \leq (1+o(1))eA_2 t^{pA_1-1} i^{-pA_1}.$$

Proof. In order to simplify calculations, we make the following substitution:

$$X(v_i,t) = \deg^-(v_i,t) + \frac{A_2}{A_1}. \tag{2.4}$$

It follows from the definition of the process that

$$X(v_i,t+1) = \begin{cases} X(v_i,t)+1, & \text{with probability } \frac{pA_1X(v_i,t)}{t} \\ X(v_i,t), & \text{otherwise.} \end{cases}$$

Finding the conditional expectation,

$$\mathbb{E}(X(v_i,t+1) \mid X(v_i,t)) = (X(v_i,t)+1)\frac{pA_1X(v_i,t)}{t} + X(v_i,t)\left(1 - \frac{pA_1X(v_i,t)}{t}\right)$$
$$= X(v_i,t)\left(1+\frac{pA_1}{t}\right).$$

Taking expectations again, we get

$$\mathbb{E}(X(v_i,t+1)) = \mathbb{E}(X(v_i,t))\left(1+\frac{pA_1}{t}\right).$$

Since all nodes start with in-degree zero, $X(v_i,i) = \frac{A_2}{A_1}$. Note that, for $0 < x < 1$, $\log(1+x) = x - O(x^2)$. If $i \gg 1$, one can use this to get

$$\mathbb{E}(X(v_i,t)) = \frac{A_2}{A_1}\prod_{j=i}^{t-1}\left(1+\frac{pA_1}{j}\right) = (1+o(1))\frac{A_2}{A_1}\exp\left(\sum_{j=i}^{t-1}\frac{pA_1}{j}\right),$$

but in all cases $i \geq 1$,

$$\mathbb{E}(X(v_i,t)) \leq \frac{A_2}{A_1}\exp\left(\sum_{j=i}^{t-1}\frac{pA_1}{j}\right).$$

Therefore, when $i \gg 1$,

$$\mathbb{E}(X(v_i,t)) = (1+o(1))\frac{A_2}{A_1}\exp\left(pA_1\log\left(\frac{t}{i}\right)\right) = (1+o(1))\frac{A_2}{A_1}\left(\frac{t}{i}\right)^{pA_1},$$

and (2.2) follows from (2.4) and (2.1). Moreover, for any $i \geq 1$

$$\mathbb{E}(X(v_i,t)) \leq \frac{A_2}{A_1}\exp\left(pA_1\left(\log\left(\frac{t}{i}\right)+1/i\right)\right) \leq \frac{eA_2}{A_1}\left(\frac{t}{i}\right)^{pA_1},$$

and (2.3) follows from (2.4) and (2.1) as before, which completes the proof.

Another fact that we will need follows directly from the following result proved in [21]. The degree of an individual vertex is not concentrated, due to variation happening shortly after birth. (That is, a.a.s. there are vertices that have smaller/larger degrees than what we would expect.) However, provided that the degree of the vertex at end time t is large enough (that is, is tending to infinity faster than $\log t$), sharp bounds on the degree of the vertex during most of the process were obtained. This is expressed in the following theorem. First, define a injective function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(i) = \frac{A_2}{A_1}\left(\frac{t}{i}\right)^{pA_1},$$

so $f(i)$ is the expected degree, at time t, of a vertex born at time i (up to a factor of $(1+o(1))$). Thus, $f^{-1}(k)$ is the birth time of a vertex of final degree k, assuming the degree of the vertex is close to the expected value during its lifetime. Hence, if a vertex of final degree k has behaviour close to its expected degree, then

$$t_a = f^{-1}\left(\frac{A_2 k}{A_1 a}\right)$$

will be the time when that vertex has degree a. Indeed, for a vertex born at time $f^{-1}(k)$, the expected degree at time t_a is equal to

$$\begin{aligned}\frac{A_2}{A_1}\left(\frac{t_a}{f^{-1}(k)}\right)^{pA_1} &= \frac{A_2}{A_1}\left(\frac{A_2}{A_1}\left(\frac{t}{f^{-1}(k)}\right)^{pA_1}\right)\Bigg/\left(\frac{A_2}{A_1}\left(\frac{t}{t_a}\right)^{pA_1}\right)\\ &= \frac{A_2 k}{A_1}\Bigg/\left(\frac{A_2 k}{A_1 a}\right) = a.\end{aligned}$$

Theorem 2 ([21]). *Let $\omega = \omega(t)$ be any function tending to infinity together with t. The following statement holds a.a.s. for every vertex v for which $\deg^-(v,t) = k = k(t) \geq \omega\log t$. Let $i = f^{-1}(k)$, and let t_k be*

$$t_k = f^{-1}\left(\frac{A_2 k}{A_1\omega\log t}\right).$$

Then, for all values of s such that $t_k \leq s \leq t$,

$$\deg^-(v,s) = (1+o(1))\frac{A_2}{A_1}\left(\frac{s}{i}\right)^{pA_1} = (1+o(1))k\left(\frac{s}{t}\right)^{pA_1}. \tag{2.5}$$

The theorem implies that once a given vertex accumulates $\omega \log t$ neighbours, the rest of the process (until time-step t) can be predicted with high probability; in fact, a.a.s. we get a concentration around the expected value.

Now, with Theorem 2 in hand, we get immediately the following.

Theorem 3. *Let $\omega = \omega(t)$ is a function that goes to infinity together with t. The following holds a.a.s. for every vertex v_i added at time i. For all $i \le s \le t$ we have*

$$\deg^-(v_i, s) = O\left((\omega \log t)\left(\frac{s}{i}\right)^{pA_1}\right),$$
$$|S(v_i, s)| = O\left(\frac{\omega \log t}{i}\right).$$

Proof. For a contradiction suppose that $k = \deg^-(v_i, s) \ge (2\omega \log t)(s/i)^{pA_1}$ for some value of s ($i \le s \le t$). Since $k \ge \omega \log t$, Theorem 2 can be applied to get that

$$\begin{aligned}
\deg^-(v_i, i) &= (1 + o(1))\frac{A_2}{A_1}\left(\frac{i}{f^{-1}(k)}\right)^{pA_1} \\
&= (1 + o(1))\frac{A_2}{A_1}\left(\frac{s}{f^{-1}(k)}\right)^{pA_1}\left(\frac{s}{i}\right)^{-pA_1} \\
&= (1 + o(1))k\left(\frac{s}{i}\right)^{-pA_1} \\
&\ge (2 + o(1))\omega \log t,
\end{aligned}$$

which is clearly a contradiction (in fact, $\deg^-(v_i, i) = 0$).

3. Directed diameter

The small world property, introduced by Watts and Strogatz [31], is a central notion in the study of complex networks (see also [24]). The small world property demands a low diameter of $O(\log t)$, and a higher clustering coefficient than found in a binomial random graph with the same number of nodes and same average degree. Adamic et al. [1] provided an early study of a social network at Stanford University, and found that the network has the small world property. Similar results were found in [2] which studied Cyworld, MySpace, and Orkut, and in [28] which examined data collected from Flickr, YouTube, LiveJournal, and Orkut. Low diameter (of 6) and high clustering coefficient were reported in the Twitter by both Java et al. [23] and Kwak et al. [25]. Many well-known models for complex networks, including the preferential attachment model by Barabási and Albert [5], have diameters growing at most logarithmically with time. (In fact, in [8] Bollobás and Riordan showed that a.a.s. the diameter of the preferential attachment model is asymptotic to $\log t / \log \log t$.)

Consider a graph G_t produced by the SPA model. For a given pair of vertices $v_i, v_j \in V_t$ ($1 \le i < j \le t$), let $l(v_i, v_j)$ denote the length of the shortest directed

path from v_j to v_i if such a path exists, and let $l(v_i, v_j) = 0$ otherwise. The directed diameter of a graph G_t is defined as

$$D(G_t) = \max_{1 \le i < j \le t} l(v_i, v_j).$$

The next subsection (Subsection 3.1) is devoted to proving the following result on the upper bound of $D(G_t)$:

Theorem 4. *Consider the SPA model. There exists an absolute constant $c_1 > 0$ such that a.a.s.*

$$D(G_t) \le c_2 \log t.$$

Analyzing the lower bound appears to be more challenging and more technical. In order to avoid some additional technicalities, we will focus on 2-dimensional Euclidean metric and will assume that some extra condition (namely, that $A_1 < 3A_2$) holds. Generalizing the result and removing the condition seems to be possible but, since it is not clear at the moment whether the upper or the lower bound (or neither) is correct, we do not do it. The proof of the following result can be found in Subsection 3.2.

Theorem 5. *Consider the SPA model for 2-dimensional Euclidean metric, and assume that $A_1 < 3A_2$. There exists an absolute constant $c_2 > 0$ such that a.a.s.*

$$D(G_t) \ge \frac{c_1 \log t}{\log \log t}.$$

3.1. Upper bound

An $O(\log t)$ upper bound on the directed diameter is obtained as follows.

Theorem 6. *Let $C = 18 \max(A_2, 1)$. With probability $1 - o(t^{-2})$ we have that for any $1 \le i < j \le t$, G_t does not contain a directed (v_i, v_j)-path of length at least $k^* = C \log t$.*

As there are at most t^2 pairs v_i, v_j, the Theorem 4 will follow as well.

Proof. In order to simplify the notation, we use v to denote the vertex added at step $v \le t$. Let vPu be a directed (v, u)-path of length given by $vPu = (v, t_{k-1}, t_{k-2}, \dots, t_1, u)$, let $t_0 = u, t_k = v$.

$$\Pr(vPu) = \prod_{i=1}^{k} p \left(\frac{A_1 \deg^-(t_{i-1}, t_i) + A_2}{t_i} \right).$$

Let $N(v, u, k)$ be the number of directed (v, u)-paths of length k, then

$$\mathbb{E}N(v, u, k) = \sum_{u < t_1 < \dots < t_{k-1} < v} p^k \mathbb{E} \left(\prod_{i=1}^{k} \left(\frac{A_1 \deg^-(t_{i-1}, t_i) + A_2}{t_i} \right) \right).$$

However

$$\mathbb{E}(\deg^-(t_i,t_{i+1}) \mid \deg^-(t_{j-1},t_j) \quad \text{and} \quad (t_{j-1},t_j)\in E_t,\, j\le i) = \mathbb{E}(\deg^-(t_i,t_{i+1})).$$

We first consider the case where u tends to infinity together with t. ¿From Theorem 1 it follows that

$$\mathbb{E}(\deg^-(t_{i-1},t_i)) = (1+o(1))\frac{A_2}{A_1}\left(\frac{t_i}{t_{i-1}}\right)^{pA_1} - \frac{A_2}{A_1}.$$

Thus

$$\begin{aligned}
\mathbb{E}N(v,u,k) &= \sum_{u<t_1<\cdots<t_{k-1}<v} p^k \prod_{i=1}^{k}\frac{1}{t_i}\left(A_1\mathbb{E}(d^-(t_{i-1},t_i))+A_2\right)\\
&= \sum_{u<t_1<\cdots<t_{k-1}<v} (1+o(1))^k (A_2p)^k \prod_{i=1}^{k}\frac{1}{t_i}\left(\frac{t_i}{t_{i-1}}\right)^{pA_1}\\
&= (1+o(1))^k (A_2p)^k \left(\frac{v}{u}\right)^{pA_1}\frac{1}{v}\sum_{u<t_1<\cdots<t_{k-1}<v}\prod_{i=1}^{k-1}\frac{1}{t_i}.
\end{aligned}$$

However

$$\begin{aligned}
\sum_{u<t_1<\cdots<t_{k-1}<v}\prod_{i=1}^{k-1}\frac{1}{t_i} &\le \frac{1}{(k-1)!}\left(\sum_{u<s<v}\frac{1}{s}\right)^{k-1}\\
&\le \frac{1}{(k-1)!}\left(\log v/u + 1/u\right)^{k-1}\\
&\le \left(\frac{e(\log v/u+1/u)}{k-1}\right)^{k-1}.
\end{aligned}$$

Let $k^* = C\log t$, where $C = 18\max(1,A_2)$. Assuming t sufficiently large, and recalling that $pA_1<1$, we have

$$\begin{aligned}
\sum_{k>k^*}\mathbb{E}N(v,u,k) &\le 2A_2\sum_{k>k^*}\left(\frac{(1+o(1))A_2pe(\log v/u+1/u)}{k-1}\right)^{k-1}\\
&\le 2A_2\left(\frac{(1+o(1))A_2e(\log v/u+1/u)}{C\log t}\right)^{k^*}\frac{1}{1-3A_2/C}\\
&= O(6^{-18\log t})\\
&= o(t^{-4}).
\end{aligned}$$

The result follows for u tending to infinity. In the case where u is a constant, it follows from Theorem 1 that a multiplicative correction of e can be used in $\mathbb{E}(\deg^-(t_{i-1},t_i))$, leading to an error term of $O(t^{-18\log 2}) = o(t^{-4})$, as before. This finishes the proof of the lower bound.

3.2. Lower bound

In this subsection, we provide an $\Omega(\log t/\log\log t)$ lower bound on the directed diameter. We use $C(u,r)$ to denote a disk of radius r centered at vertex u, where $C(u,r(t))$ and $S(u,t)$ are related through the equations above. Let $C_a(u,r)$ denote a cap of area a relative to the area πr^2 of $C(u,r)$. To form the cap of the disk $C(u_0,r)$ centred at $u_0=(0,0)\in\mathbb{R}^2$ we take the points $\{(x,y):\rho r\le x\le r, x^2+y^2\le r\}\subseteq C(u_0,r)$. Here $0<\rho<1$ is taken to be an absolute constant sufficiently close to one to make some claimed inequalities below valid. The absolute area $\widehat{a}(\rho)$ of this cap is given by $\widehat{a}(\rho)=r^2(\pi/2-\rho\sqrt{1-\rho^2}-\sin^{-1}\rho)$. We note that the relative area $a=\widehat{a}/\pi r^2$ of the cap is not a function of r.

Construction of a good sequence of disks

We use the notation $r=r(t)=\sqrt{A_2/\pi t}$ and $r'=r'(t)=\sqrt{(A_1+A_2)/\pi t}$, to indicate the radius of disks (at time t) with vertices of in-degree zero and of in-degree one, respectively. The condition that $r'<2r$ (used below), is equivalent to

$$\sqrt{A_1+A_2}<2\sqrt{A_2}, \tag{3.1}$$

which is equivalent to $A_1<3A_2$.

As before, in order to simplify the notation, we use v to denote the vertex added at step $v\le t$. An important condition in our construction is that if at step v a vertex v falls in $C_a(u,r(v))$ then $C(u,r'(v))\cap C_a(v,r(v))=\emptyset$. Thus if v attaches to u, so that $\deg^-(u,v)=1$, there is still a cap of $C(v,r(v))$ (namely, $C_a(v,r(v))$) that u does not reach. This condition holds provided (3.1) is true. In this way, we can construct a series of events

$$u_1\in C_a(u_0,r(u_1)),\ u_2\in C_a(u_1,r(u_2)),\cdots,u_k\in C_a(u_{k-1},r(u_k)). \tag{3.2}$$

Our construction further requires that no vertex v falls inside $C(u_0,r(v))$ at any steps $u_0<v<u_1$ or within $C(u_0,r'(v))$ and $u_1<v<t$, and the same for each u_j, $1\le j\le k$. As a consequence, $\deg^-(u_j,t)=1$ for $0\le j\le k-1$. In this way the areas of the disks are controlled at all times. Furthermore, under these circumstances, the path $u_k,u_{k-1},\dots,u_0$ will be a shortest path from u_k to u_0.

The next part of the construction is as follows. At step s we divide the unit square into horizontal strips $R(1),R(2),\dots,R(M)$ of height h and width w. Here

$$M=1/wh \quad\text{and}\quad h=4r \quad\text{and}\quad w=4(k{+}1)r \quad\text{and}\quad r=r(s)=\sqrt{A_2/\pi s}.$$

Inside a strip $R=R(i)$, there is centered a strip $R'=R'(i)$ of height $2r$ and width $(4k{+}2)r$, thus placing a boundary of depth r around R' inside R. Note that the area of R is by a factor of $(2+o(1))$ larger than the one of R' (provided that $k\to\infty$). The purpose of this construction is that any disk of radius r centered in R' must be contained within R. Therefore, if two paths, $u_k^1,u_{k-1}^1,\dots,u_0^1$ and $u_k^2,u_{k-1}^2,\dots,u_0^2$, are constructed such that $u_j^i\in R'(i)$, $i=1,2$, $j=0,1,\dots,k$, then the events corresponding to the two strips are independent. Moreover, if $u\in R'(i)$, then at least half of the cup $C_a(u,r)$ falls in $R'(i)$.

We will argue that a.a.s. at least one strip will contain a sequence with $k = \Omega(\log t/\log\log t)$ satisfying (3.2). We will choose $\ell = 2(k+1)$ where $k > 0$ is integer, and $s = t/\log t$. The rectangle of size $L = 2r\ell$ is used to initialize the process, using some point $u = u_0$; and the sequence of k squares of side $2r$ will be enough to contain the subsequent vertices in the construction (3.2) above.

Probability estimates for good sequences

We suppose that the construction of a good sequence of disks occurs in some R and are centered in R'. Given some set of steps $s < u_0 < u_1 < \cdots < u_k < t$, let $\mathcal{E}(u_0, u_1, \ldots, u_k)$ be the event that the construction occurred at these steps and that u_j attaches to u_{j-1}, $j = 1, 2, \ldots, k$. This forms a directed path from u_k to u_0 with the property that there are no short cuts, i.e. no u_j attaches to any u_i where $i < j - 1$.

$$\begin{aligned}
&\Pr(\mathcal{E}(u_0, u_1, \ldots, u_k)) \\
&\quad \geq \Pr(u_0 \in R') \prod_{\tau=u_0+1}^{u_1-1} \left(1 - \frac{A_2}{\tau}\right) \frac{p(a/2)A_2}{u_1} \\
&\qquad\qquad \prod_{\tau=u_1+1}^{u_2-1} \left(1 - \frac{A_1 + 2A_2}{\tau}\right) \frac{p(a/2)A_2}{u_2} \cdots \\
&\qquad\qquad \cdots \prod_{\tau=u_{k-1}+1}^{u_k-1} \left(1 - \frac{(k-2)A_1 + (k-1)A_2}{\tau}\right) \frac{p(a/2)A_2}{u_k} \\
&\qquad\qquad \prod_{\tau=u_k+1}^{t} \left(1 - \frac{(k-1)A_1 + kA_2}{\tau}\right).
\end{aligned}$$

Let $q = A_1 + A_2$. If $x = o(1)$ then $1 - x = e^{-x - O(x^2)}$, and so, assuming $s \to \infty$,

$$\begin{aligned}
&\Pr(\mathcal{E}(u_0, u_1, \ldots, u_k)) \\
&\quad \geq \frac{1}{(2+o(1))M} (apA_2/2)^k \frac{1}{u_1} \cdots \frac{1}{u_k} \\
&\qquad \times \exp\left\{ -\sum_{i=0}^{k} (iA_1 + (i+1)A_2) \sum_{\tau=u_i+1}^{u_{i+1}} \left(\frac{1}{\tau} + O\left(\frac{i}{\tau^2}\right)\right)\right\} \\
&\quad \geq \frac{1}{3M} e^{-O(tk^2/s^2)} (apA_2/2)^k \frac{1}{u_1} \cdots \frac{1}{u_k} \left(\frac{u_0}{u_1}\right)^{A_2} \left(\frac{u_1}{u_2}\right)^{A_1+2A_2} \cdots \left(\frac{u_k}{t}\right)^{(k-1)A_1+kA_2} \\
&\quad \geq \frac{1}{4M} (apA_2/2)^k \frac{u_0^{A_2} t^{A_1}}{t^{kq}}\ {u_1}^{q-1} {u_2}^{q-1} \cdots {u_k}^{q-1},
\end{aligned}$$

where the last line assumes that $tk^2 = o(s^2)$.

We note that

$$\sum_{s\le u_0<u_1<\cdots<u_k\le t} f(u_1)f(u_2)\cdots f(u_k) \tag{3.3}$$
$$\ge \frac{1}{k!}\left(\sum_{\tau=s}^{t} f(\tau)\right)^k - \binom{k}{2}\frac{1}{k-2!}\left(\sum_{\tau} f^2(\tau)\right)\left(\sum_{\tau} f(\tau)\right)^{k-2}.$$

Thus

$$\sum_{s\le u_0<u_1<\cdots<u_k\le t}\prod_{j=1}^{k} {u_j}^{q-1}$$
$$\ge (1+o(1))\frac{1}{k!}\left(\left(\frac{1}{q}(t^q-u_0^q)\right)^k - O(k^4)\Psi(u_0,t)\left(\frac{1}{q}(t^q-u_0^q)\right)^{k-2}\right),$$

where

$$\Psi(u_0,t)=\begin{cases}\frac{1}{1-2q}\left(\frac{1}{{u_0}^{1-2q}}-\frac{1}{t^{1-2q}}\right) & \text{if} \quad 2q<1\\ \log t-\log u_0 & \text{if} \quad 2q=1\\ \frac{1}{2q-1}\left(t^{2q-1}-{u_0}^{2q-1}\right) & \text{if} \quad 2q>1\end{cases}.$$

Let

$$\mathcal{E}(s,u_0,t)=\bigcup_{s<u_0<u_1<\cdots<u_k<t}\mathcal{E}(u_0,u_1,\cdots,u_k).$$

¿From the above it follows that

$$\Pr(\mathcal{E}(s,u_0,t))\ge\frac{1}{5M}\frac{(apA_2/2)^k}{q^k k!}\frac{u_0^{A_2}t^{A_1}}{t^{kq}}\left((t^q-u_0^q)^k-O(k^4)\Psi(u_0,t)\,(t^q-u_0^q)^{k-2}\right).$$

Provided $s=o(t)$,

$$\int_s^t z^\alpha(1-z^q)^k dz=\frac{1}{q}\int_{s/t}^1 y^{1/q+\alpha/q-1}(1-y)^k dy$$
$$=(1+o(1))\frac{1}{q}\frac{\Gamma((\alpha+1)/q)\Gamma(k+1)}{\Gamma(k+(\alpha+1)/q+1)}\ge ck^{-(\alpha+1)/q},$$

for some absolute constant $c>0$.

Let $\mathcal{E}(s,t)=\bigcup_{s<u_0<t}\mathcal{E}(s,u_0,t)$, then

$$\Pr(\mathcal{E}(s,t))\ge\frac{c}{6Mk!}\left(\frac{apA_2}{2q}\right)^k\frac{t^q}{k^{(A_2+1)/q}}.$$

Thus $\mathcal{E}(s,t)$ is the event that the construction of an isolated directed path of length k *succeeds* in a particular strip R.

Let $k=\beta\log t/\log\log t$, and let $s=t/\log t$, so that $s/t\to 0$ and $tk^2/s^2\to 0$ as required. The expected number $N(s,t)$ of strips where our construction succeeds is

$$\mathbb{E}[N(s,t)]\ge\frac{ct^q}{6k!k^{(A_2+1)/q}}\left(\frac{apA_2}{2q}\right)^k=t^{q-\beta-o(1)}.$$

Suppose first that $q = A_1 + A_2 > 1/2$. As long as $\beta < q - 1/2$, $\mathbb{E}[N(s,t)] = \Omega(t^{1/2+\epsilon})$ for some $\epsilon > 0$. The concentration of $N(s,t)$ follows from a standard martingale argument. All positions of the points v, $1 \le v \le t$ in the unit square are equally likely. Changing the location of a given point alters the value of $N(s,t)$ by at most 2. So,

$$\Pr(N(s,t) = 0) \le \exp\left\{-\Omega\left(\frac{(t^{1/2+\epsilon})^2}{t}\right)\right\} = o(1),$$

and the proof of the lower bound is complete.

Suppose now that $q = A_1 + A_2 \le 1/2$. In this case the argument for a concentration of $N(s,t)$ is slightly more technical but standard as well. Unfortunately, since the events corresponding to horizontal strips $R(1), R(2), \ldots, R(M)$ are not independent, we cannot use the Chernoff bound to get the result. However, the main effect on conditioning on a given strip (that yields a path $(w_1, w_2, \ldots, w_k)$) is to make the sum (3.3) slightly smaller, by not allowing that, for every i, u_i to be equal to w_j for some j. This reduces the sum by an amount of order at most $kt^{q(k-1)}/k!$, which is negligible comparing to the expression without the deletions. (Other effects are in our favour). In particular, for any two disjoint vectors $(u_0, u_1, \ldots, u_k)$ and $(w_0, w_1, \ldots, w_k)$, the events $\mathcal{E}(u_0, u_1, \ldots, u_k)$ and $\mathcal{E}'(w_0, w_1, \ldots, w_k)$ that correspond to different strips, we have

$$\begin{aligned}\Pr(&\mathcal{E}(u_0,u_1, \ldots, u_k) \wedge \mathcal{E}'(w_0, w_1, \ldots, w_k)) \\ &= (1 + o(1)) \Pr(\mathcal{E}(u_0, u_1, \ldots, u_k)) \Pr(\mathcal{E}'(w_0, w_1, \ldots, w_k)).\end{aligned}$$

It follows that $\Pr(\mathcal{E}(s,t) \wedge \mathcal{E}'(s,t)) = (1 + o(1)) \Pr(\mathcal{E}(s,t))^2$, $\mathrm{Var}[N(s,t)] = o(\mathbb{E}[N(s,t)]^2)$, and the concentration follows by Chebyshev's inequality.

4. Small separators

Let us note that there are some significant differences between graphs generated by the preferential attachment model and those found in the real world. One major difference is found in their expansion properties. Mihail, Papadimitriou, and Saberi [27] showed that a.a.s. the preferential attachment model has conductance bounded below by a constant. On the other hand, Blandford, Blelloch and Kash [6] found that some WWW related graphs have smaller separators than the preferential attachment model predicts. This observation is consistent with observations due to Estrada [15], who found that half of the real-world networks he looked at were good expanders and the other half were not so good. In this subsection, we show that the SPA model has small separators.

Let us recall that $V_t \subseteq S$ where S is the unit hypercube $[0,1]^m$. We use the geometry of the model to obtain a sparse cut. Let

$$S' = \left\{ s = (s_1, s_2, \ldots, s_m) \in S : s_1 < \frac{1}{2} \right\}.$$

Let us partition the vertex set V_t as follows: $V_t' = V_t \cap S'$, $V_t'' = V_t \cap (S \setminus S') = V_t \setminus V_t'$. The next theorem shows that this partition yields a sparse cut.

Theorem 7. *A.a.s. the following holds* $|V_t'| = (1+o(1))t/2$, $|V_t''| = (1+o(1))t/2$, *and*

$$|E(V_t', V_t'')| = O(t^{\max\{1-1/m, pA_1\}} \log^5 t) = o(t).$$

Proof. Clearly, we expect $t/2$ vertices in each set V_t' and V_t''. The concentration follows immediately from the Chernoff bound. It remains to show that an upper bound for the size of the cut holds a.a.s.

It follows from Theorem 3 (by taking $\omega = \log t$) that a.a.s. for every $i \in [t]$ the maximum sphere of influence of a vertex v_i added at time i is $O(i^{-1} \log^2 t)$ (during the whole process). Since we aim for a result that holds a.a.s., we may assume that this property holds for all i. Therefore, the maximum radius of influence of v_i is $O((\log^2 t/i)^{1/m})$.

We will investigate how many edges are in the cut by counting (independently) edges in this cut directed to vertices of similar age. For a given integer k such that $0 \le k < \log t$, let

$$\begin{aligned} V^{(k)} &= \{v_i \in V_t : e^k \le i < \min\{e^{k+1}, t\}\}, \\ E^{(k)} &= \{(v_i, v_j) \in E_t : v_i \in V^{(k)} \quad \text{and} \quad i < j \le t\} \\ C^{(k)} &= E^{(k)} \cap E(V_t', V_t''). \end{aligned}$$

It is clear that $\{E^{(k)} : 0 \le k < \log t\}$ is a partition of the edge set and so $\{C^{(k)} : 0 \le k < \log t\}$ is a partition of the cut $E(V_t', V_t'')$. It remains to estimate the size of $C^{(k)}$ for a given value of k.

Fix $0 \le k < \log t$, and let $v_i \in V^{(k)}$. Note that the maximum radius of influence of v_i is $O((e^{-k} \log^2 t)^{1/m})$. Therefore, if there is an edge in the cut directed to $v_i = (s_1, s_2, \ldots, s_m)$, then v_i must fall into a strip within distance $O((e^{-k} \log^2 t)^{1/m})$ from the cutting hyperplane; that is, $|s_1 - 1/2| = O((e^{-k} \log^2 t)^{1/m})$. Since $|V^{(k)}| = O(e^k)$, we get that

$$O((e^{-k} \log^2 t)^{1/m}) \cdot |V^{(k)}| = O(e^{k(1-1/m)} (\log t)^{2/m})$$

vertices of $V^{(k)}$ are expected to appear in this strip during the whole process. Hence, it follows from the Chernoff bound that with probability at least $1 - \exp(-\Theta(\log^2 t))$ there are $O(e^{k(1-1/m)} \log^2 t)$ vertices in this strip at the end of the process. (Note that the exponent of $\log t$ has changed from $2/m$ to 2 in order to guarantee the value at least $\log^2 t$ which is required for a bound to hold with the desired probability.) By Theorem 3 (again, by taking $\omega = \log t$), a.a.s. all vertices introduced in this time period have (final) in-degree at most $(t/e^k)^{pA_1} \log^2 t$, we get that

$$|C^{(k)}| = O(e^{k(1-1/m)} \log^2 t) \cdot (t/e^k)^{pA_1} \log^2 t = O(t^{pA_1} e^{k(1-1/m-pA_1)} \log^4 t)$$

edges in the cut a.a.s.

Finally, we get that a.a.s.

$$|E(V_t', V_t'')| = \sum_{k=0}^{\lceil \log t \rceil - 1} |C^{(k)}| = \sum_{k=0}^{\lceil \log t \rceil - 1} O(t^{pA_1} e^{k(1-1/m-pA_1)} \log^4 t)$$
$$\leq \begin{cases} \log t \cdot O(t^{pA_1} t^{1-1/m-pA_1} \log^4 t) = O(t^{1-1/m} \log^5 t), & \text{if } pA_1 < 1 - 1/m; \\ \log t \cdot O(t^{pA_1} \log^4 t) = O(t^{pA_1} \log^5 t), & \text{otherwise,} \end{cases}$$

which finishes the proof.

As we already mentioned, it is believed that a large fraction of real-world networks possess bad spectral expansion properties realized by relatively large gaps between the first and second eigenvalues of their adjacency matrices. The fact that the SPA model has sparse cuts easily implies bad spectral expansion properties.

The normalized Laplacian of a graph relates to important graph properties; see [12]. Let A denote the adjacency matrix and D denote the diagonal degree matrix of a graph G. Then the normalized Laplacian of G is $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$. Let $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1} \leq 2$ denote the eigenvalues of $\mathcal{L}$. The *spectral gap* of the normalized Laplacian is

$$\lambda = \max\{|\lambda_1 - 1|, |\lambda_{n-1} - 1|\}.$$

A spectral gap bounded away from zero is an indication of bad expansion properties. The next theorem represents a drastic departure from the good expansion found in binomial random graphs, where $\lambda = o(1)$ [12, 13].

Theorem 8. *Consider the SPA model. Let $\lambda = \lambda(t)$ be the spectral gap of the normalized Laplacian of G_t. Then a.a.s.*

$$\lambda(t) = 1 + o(1).$$

In order to prove this result, we use the expander mixing lemma for the normalized Laplacian (see [12] for its proof). For two sets of vertices X and Y we use the notation $\mathrm{vol}(X)$ for the volume of the subgraph induced by X, $\bar{X}$ for the complement of X, and $e(X, Y)$ for the number of edges with one end in each of X and Y. (Note that $X \cap Y$ does not have to be empty; in general, $e(X, Y)$ is defined to be the number of edges between $X \setminus Y$ to Y plus twice the number of edges that contain only vertices of $X \cap Y$.)

Lemma 1. *Let λ be the spectral gap of the normalized Laplacian of G. For all sets $X \subseteq G$,*

$$\left| e(X, X) - \frac{(\mathrm{vol}(X))^2}{\mathrm{vol}(G)} \right| \leq \lambda \frac{\mathrm{vol}(X)\mathrm{vol}(\bar{X})}{\mathrm{vol}(G)}.$$

Now, we are ready to come back to the proof of Theorem 8.

Proof (Proof of Theorem 8). Using the notation introduced before Theorem 7 and the theorem itself we get that a.a.s.

$$\begin{aligned} \mathrm{vol}(V_t') = (1+o(1))\mathrm{vol}(\bar{V}_t') &= \Theta(t) \\ \mathrm{vol}(G) = (2+o(1))\mathrm{vol}(V_t') & \\ e(V_t', V_t') = \mathrm{vol}(V_t') - e(V_t', V_t'') &= (1+o(1))\mathrm{vol}(V_t'). \end{aligned}$$

It follows from Lemma 1 (applied to $X = V_t'$) that a.a.s. $\lambda(t) \geq 1 + o(1)$. By definition, $\lambda(t) \leq 1$ so $\lambda(t) = 1 + o(1)$.

5. Emergence of giant component

Let us note that all edges in G_t are from younger vertices to older ones; that is, denoting by v_i the vertex added at time i we get that if $(v_j, v_i) \in E_t$, then $j > i$. This implies that G_t has t strongly connected components, each of which consists of one vertex.

On the other hand, it seems that investigating the size of the largest weak connected component is a non-trivial task. Let $\hat{G}_t = (V_t, \hat{E}_t)$ be the underlying graph of G_t; that is, $\hat{G}$ is an undirected graph on the vertex set V_t and $\{v_j, v_i\} \in \hat{E}_t$ if and only if $(v_j, v_i) \in E_t$. We wish to know the size of the largest component in $\hat{G}_t$.

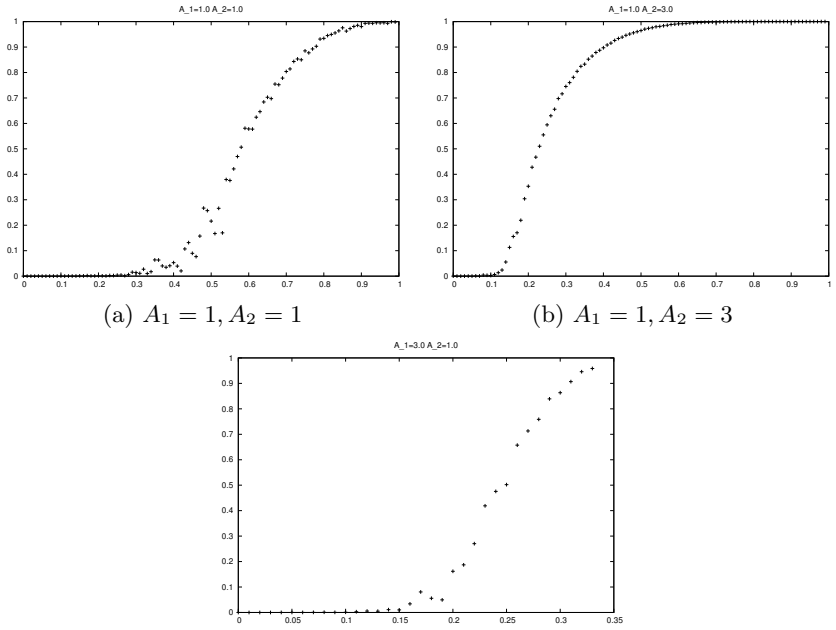
One can show that the expected number of edges added at time t of the process is $\deg^+(v_t, t) = \frac{pA_2}{1-pA_1}$. Therefore, if $p > p_1 := (A_1 + A_2)^{-1}$, then the expected out degree in G_t is larger than 1, and so is the expected degree in $\hat{G}_t$. By looking at the 'branching factor' of the breadth-first search process it is natural to conjecture that a.a.s. there exists a giant component if $p > p_1$. On the other hand, if $p < p_1$, then the expected out-degree in G_t is smaller than one, but this fact does not in itself guarantee absence of the giant component in $\hat{G}_t$. Is p_1 the threshold we search for? If $p < p_2 := (A_1 + 2A_2)^{-1}$, then $\deg^+(v_t, t) < 1/2$ and so the average degree in $\hat{G}_t$ is smaller than one. Perhaps p_2 is the threshold for the giant component? Clearly, more sophisticated argument is required to solve this problem and we will try to settle this down in the journal version of this paper.

We performed a number of simulations to make a better prediction (see Figure 2). For a given set of parameters A_1, A_2, we performed a number of simulations ($p = i/100$, $0 \leq i < 1/A_1$). Unfortunately, it seems that $t = 100,000$ is still too small to observe a clear trend. However, based on these numerical results, one can conjecture the following.

Conjecture 1. $p_3 := (2A_1 + 2A_2)^{-1}$ is the threshold for the giant component.

References

1. L.A. Adamic, O. Buyukkokten, and E. Adar, A social network caught in the web, *First Monday* **8** (2003).



(a) $A_1 = 1, A_2 = 1$ (b) $A_1 = 1, A_2 = 3$

(c) $A_1 = 3, A_2 = 1$

Fig. 2. A simulation of the SPA model on the unit 2-dimensional torus with $t = 100,000$. (The x-axis is p, y-axis is the fraction of vertices in the largest component of $\hat{G}_t$.)

2. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, Analysis of topological characteristics of huge on-line social networking services, In: *Proceedings of the 16th International Conference on World Wide Web*, 2007.
3. W. Aiello, A. Bonato, C. Cooper, J. Janssen, and P. Pralat, A spatial web graph model with local influence regions, *Internet Mathematics* **5** (2009), 175–196.
4. W. Aiello, A. Bonato, C. Cooper, J. Janssen, and P. Pralat, A spatial web graph model with local influence regions, In: *Proceedings of the 5th Workshop on Algorithms and Models for the Web-Graph (WAW2007)*, Lecture Notes in Computer Science **4863**, Springer, 2007, 96–107.
5. A. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999), 509–512.
6. D. Blandford, G.E. Blelloch, and I. Kash, Compact Representations of Separable Graphs, In *Proc. of ACM/SIAM Symposium on Discrete Algorithms* (2003) 679–688.
7. B. Bollobás and O. Riordan, Mathematical results on scale-free graphs, *Handbook of Graphs and Networks*, S. Bornholdt, H. Schuster (eds), Wiley-VCH, Berlin (2002).
8. B. Bollobás and O. Riordan, The diameter of a scale-free random graph, *Combinatorica*, **4** (2004) 5–34.
9. A. Bonato, *A course on the web graph*, American Mathematical Society Graduate Studies in Mathematics 2008, Volume 89, (2008).

10. A. Bonato, J. Janssen, and P. Pralat, Geometric Protean Graphs, *Internet Mathematics* **8** (2012), 2–28.
11. M. Bradonjic, A. Hagberg, and A.G. Percus, The structure of geographical threshold graphs, *Internet Mathematics* **4** (2009), 113–139.
12. F.R.K. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, Rhode Island, 1997.
13. F.R.K. Chung, L. Lu, *Complex Graphs and Networks*, American Mathematical Society, 2006.
14. C. Cooper, A. Frieze, and P. Pralat, Some typical properties of the Spatial Preferred Attachment model, Proceedings of the 9th Workshop on Algorithms and Models for the Web Graph (WAW 2012), Lecture Notes in Computer Science **7323**, Springer, 2012, 29–40.
15. E. Estrada, Spectral scaling and good expansion properties in complex networks, *Europhysics Letters* **73** (4) (2006) 649–655.
16. M. Faloutsos, P. Faloutsos, and C. Faloutsos, On Power-law Relationships of the Internet Topology, *SIGCOMM* (1999) 251–262.
17. A. Flaxman, A.M. Frieze, and J. Vera, A geometric preferential attachment model of networks, *Internet Mathematics*, 3(2):187–206, 2006.
18. A. Flaxman, A.M. Frieze, and J. Vera, A geometric preferential attachment model of networks II, *Internet Mathematics*, 4(1):87–111, 2008.
19. D.J. Higham, M. Rasajski, and N. Przulj, Fitting a geometric graph to a protein-protein interaction network, *Bioinformatics*, 24(8):1093–1099, 2008.
20. J. Janssen, Spatial models for virtual networks, In: *Programs, Proofs, Processes: 6th international conference on Computability in Europe (CiE10)*, Ferreira *et al.* (eds.), Springer LNCS 6158 (2010), pp. 201–210.
21. J. Janssen, P. Pralat, and R. Wilson, Geometric Graph Properties of the Spatial Preferred Attachment model, accepted to *Advances in Applied Mathematics*.
22. J. Janssen, P. Pralat, and R. Wilson, Estimating node similarity from co-citation in a spatial graph model, In: *Proceedings of the 2010 ACM Symposium on Applied Computing–Special Track on Self-organizing Complex Systems*, 2010, 1329–1333.
23. A. Java, X. Song, T. Finin, and B. Tseng, Why we twitter: understanding microblogging usage and communities, In: *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, 2007.
24. J. Kleinberg, The small-world phenomenon: An algorithmic perspective, In: *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.
25. H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media?, In: *Proceedings of the 19th International World Wide Web Conference*, 2010.
26. N. Masuda, M. Miwa, and N. Konno, Geographical threshold graphs with small-world and scale-free properties, *Phys. Rev. E*, 71(3):036108, 2005.
27. M. Mihail, C.H. Papadimitriou, and A. Saberi, On Certain Connectivity Properties of the Internet Topology, In *Proc. IEEE Symposium on Foundations of Computer Science* (2003) 28.
28. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, Measurement and analysis of on-line social networks, In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.
29. M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, In: *Proc. of the 39th Annual Allerton Conf. on Communication, Control, and Computing* (2001), 182–191.

30. G. Yule, A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, *Philosophical Transactions of the Royal Society of London (Series B)* **213** (1924), 21–87.
31. D.J. Watts and S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440–442.