A spatial web graph model with local influence regions

W. Aiello¹, A. Bonato², C. Cooper³, J. Janssen⁴, and P. Prałat⁴

¹ University of British Columbia
   Vancouver, Canada
   aiello@cs.ubc.ca
² Wilfrid Laurier University
   Waterloo, Canada
   abonato@rogers.com
³ King's College
   London, UK
   colin.cooper@kcl.ac.uk
⁴ Dalhousie University
   Halifax, Canada
   janssen@mathstat.dal.ca, pralat@mathstat.dal.ca

Abstract. The web graph may be considered as embedded in a topic space, with a metric that expresses the extent to which web pages are related to each other. Using this assumption, we present a new model for the web and other complex networks, based on a spatial embedding of the nodes, called the Spatial Preferred Attachment (SPA) model. In the SPA model, nodes have influence regions of varying size, and new nodes may only link to a node if they fall within its influence region. We prove that our model gives a power law in-degree distribution, with exponent in [2, ∞) depending on the parameters, and with concentration for a wide range of in-degree values. We also show that the model allows for edges that span a large distance in the underlying space, modelling a feature often observed in real-world complex networks.

1 Introduction

Current stochastic models for complex networks (such as those described in [1, 2]) aim to reproduce a number of graph properties observed in real-world networks such as the web graph. On the other hand, experimental and heuristic treatments of real-life networks operate under the tacit assumption that the network is a visible manifestation of an underlying hidden reality. For example, it is commonly assumed that communities in a social network can be recognized as densely linked subgraphs, or that web pages with many common neighbours contain related topics. Such assumptions imply that there is an a priori community structure or relatedness measure of the nodes, which is reflected by the link structure of the graph.

A common method to represent relatedness of objects is by an embedding in a metric space, so that related objects are placed close together, and communities are represented by clusters of points. Following a common text mining technique, web pages are often represented as vectors in a word-document space. Using Latent Semantic Indexing, these vectors can then be embedded in a Euclidean topic space, so that pages on similar topics

⋆ The authors gratefully acknowledge support from NSERC and MITACS grants.
are located close together. Experimental studies [7] have confirmed that similar pages are more likely to link to each other. On the other hand, experiments also confirm a large amount of topic drift: it is possible to move to a completely different topic in a relatively short number of hops. This points to a model where nodes are embedded in a metric space, and the edge probability between nodes is influenced by their proximity, but edges that span a larger distance in the space are not uncommon.

The Spatial Preferred Attachment (SPA) model proposed in this paper combines the above considerations with the often-used preferential attachment principle: pages with high in-degree are more likely to receive new links. In the SPA model, each node is placed in space and surrounded by an influence region. The area of the influence region is determined by the in-degree of the node. Moreover, in each time-step all regions decrease in area as a function of time. A new node $v$ can only link to an existing node $u$ if $v$ falls within the influence region of $u$. If $v$ falls within the region of influence $u$, then $v$ will link to $u$ with probability $p$. Thus, the model is based on the preferential attachment principle, but only implicitly: nodes with high in-degree have a large region of influence, and therefore are more likely to attract new links.

A random graph model with certain similarities to the SPA model is the geometric random graph; see [8]. In that model, all influence regions have the same size, and the link probability is $p = 1$. Flaxman, Frieze, and Vera in [5] supply an interesting geometric model where nodes are embedded on a sphere, and the link probability is influenced by the relative positions of the nodes. This model is a generalization of a geometric preferential attachment models presented by the same authors in [4], which influenced our model.

There are at least three features that distinguish the SPA model from previous work. First, a new node can choose its links purely based on local information. Namely, the influence region of a node can be seen as the region where a web page is visible: only web pages that are close enough (in topic) to fall within the influence region will be aware of the give page, and thus have a possibility to link to it. Moreover, a new node links independently to each node visible to it. Consequently, the new node needs no knowledge of the invisible part of the graph (such as in-degree of other nodes, or total number of nodes or links) to determine its neighbourhood. Second, since a new node links to each visible node independently, the out-degree is not a constant nor chosen according to a pre-determined distribution, but arises naturally from the model. Third, the varying size of the influence regions allows for the occasional long links, edges between nodes that are spaced far apart. This implies a certain ”small world” property.

We formally define the SPA model as follows. Let $S$ be the surface of the sphere of area 1 in $\mathbb{R}^3$. For each positive real number $\alpha \leq 1$, and $u \in S$, define the cap around $u$ with area $\alpha$ as

$$B_\alpha(u) = \{ x \in S : ||x - u|| \leq r_\alpha \},$$

where $|| \cdot ||$ is the usual Euclidean norm, and $r_\alpha$ is chosen such that $B_\alpha$ has area $\alpha$.

The SPA model has parameters $A_1, A_2, A_3, p \geq 0$ such that $p \leq 1$, $A_1 \leq 1$ and $A_2 > 0$. It generates stochastic sequences of graphs $(G_t : t \geq 0)$, where $G_t = (V_t, E_t)$,
and \( V_t \subseteq S \). Let \( d^-(v,t) \) (\( d^+(v,t) \)) be the in-degree (out-degree) of node \( v \) in \( G_t \). We define the influence region of node \( v \) at time \( t \geq 1 \), written \( R(v,t) \), to be the cap around \( v \) with area

\[
|R(v,t)| = \frac{A_1 d^-(v,t) + A_2}{t + A_3},
\]

or \( R(v,t) = S \) if the right-hand-side is greater than 1.

The process begins at \( t = 0 \), with \( G_0 \) being the empty 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 node \( v_t \) is chosen uniformly at random (uar) from \( S \), and added to \( V_{t-1} \) to create \( V_t \). Next, independently, for each node \( u \in V_{t-1} \) such that \( v_t \in R(u,t-1) \), a directed edge \( (v_t,u) \) is created with probability \( p \). Thus, the probability that a link \( (v_t,u) \) is added in time-step \( t \) equals \( p|R(u,t-1)| \).

Because new nodes choose independently whether to link to each visible node, and the size of the influence region of a node depends only on the edges from younger nodes, the distribution of the random graph \( G_n \) produced by the SPA model with parameters \( A_1, A_2, A_3, p \) is equivalent to the graph \( G_{n+A_3} \) produced by the SPA model with the same values for \( A_1, A_2, p \), but with \( A_3 = 0 \), where the first \( A_3 \) nodes have been removed. Since the results presented in this paper do not depend on the first nodes, we will assume throughout that \( A_3 = 0 \).

Note that the model could be defined on any compact set of measure 1. However, if the set has non-empty boundary, the definition of the influence regions should be adjusted. If higher dimensions are desired, \( S \) could be chosen to be the boundary of a hypersphere in \( \mathbb{R}^k \) for some \( k \). The results in Sections 2 and 3 will still hold, while Section 4 can be easily extended to this case.

We prove in Section 2 that with high probability a graph \( G_n \) generated by the SPA model has an in-degree distribution that follows a power law in-degree distribution with exponent \( 1 + \frac{1}{p A_1} \), with concentration up to \( n^{i_f} \), where \( i_f = \left( \frac{n}{\log^2 n} \right)^{A_1/(6pA_1+2)} \). If \( p A_1 = 10/11 \), then the power law in-degree exponent is 2.1, the same as observed in the web graph (see, for example [2]). We also give a precise expression for the probability distribution of each individual node \( v_t \), provided that \( p A_1 < 1 \). In Section 3, we show that, if \( p A_1 < 1 \), the number of edges of \( G_n \) is linear, and strongly concentrated around the mean, while if \( p A_1 = 1 \) the expected number of edges is \( n \log n \). In Section 4 we explore a geometric version of the small world property. We show that the expected sum of (geometric) lengths of new edges added at time \( t \) in the SPA model is \( \Theta(t^{2-b}) \), where \( b = 1 + \frac{1}{p A_1} \) is the exponent of the power law. For the in-degree power law exponent \( b = 2.1 \) commonly observed in the web graph, this expected sum of lengths is greater than the corresponding expected sum in a corresponding geometric random graph with equal-sized influence regions.
2 In-degree distribution

In the rest of the paper, \((G_t : t \geq 0)\) refers to a sequence of random graphs generated by
the SPA model with parameters \(A_1, A_2, A_3 = 0\), and \(p\). In this section, we explore the
in-degree of the nodes in \(G_n\). We say that an event holds asymptotically almost surely
\((aas)\) if it holds with probability tending to one as \(n \to \infty\); an event holds with extreme
probability \((wep)\) if it holds with probability at least \(1 - \exp(-\Theta(\log^2 n))\) as \(n \to \infty\).
Let \(N_{i,t}\) denote the number of nodes of in-degree \(i\) in \(G_t\). For an integer \(n \geq 0\), define
\[
i_f = f(n) = \left(\frac{n}{\log^4 n}\right)^{pA_1/(6pA_1 + 2)}.
\]
(1)

Our main result in this section is the following.

Theorem 1. Fix \(p \in (0, 1]\). Then for any \(i \geq 0\),
\[
\mathbb{E}(N_{i,n}) = c_i n(1 + o(1)),
\]
(2)
where
\[
c_0 = \frac{1}{1 + pA_2},
\]
(3)
and for \(1 \leq i \leq n\),
\[
c_i = \frac{p^i}{1 + pA_2 + ipA_1} \prod_{j=0}^{i-1} \frac{jA_1 + A_2}{1 + pA_2 + jpA_1}.
\]
(4)
For \(i = 0, \ldots, i_f\), \(wep\)
\[
N_{i,n} = c_i n(1 + o(1)).
\]
(5)

Since \(c_i = ci^{-(1+\frac{1}{pA_1})}(1 + o(1))\) for some constant \(c\), this shows that for large \(i\), the
expected proportion \(N_{i,n}/n\) follows a power law with exponent \(1 + \frac{1}{pA_1}\), with concen-
tration for all values of \(i\) up to \(i_f\). The proof of the Theorem 1 is contained in the rest
of this section.

2.1 Expected value

The equations relating the random variables \(N_{i,t}\) are described as follows. As \(G_1\) consist
of one isolated node, \(N_{0,1} = 1\), and \(N_{i,1} = 0\) for \(i > 0\). For all \(t > 0\), we derive that
\[
\mathbb{E}(N_{0,t+1} - N_{0,t} \mid G_t) = 1 - N_{0,t} p \frac{A_2}{t},
\]
(6)
\[
\mathbb{E}(N_{i,t+1} - N_{i,t} \mid G_t) = N_{i-1,t} p \frac{A_1(i - 1) + A_2}{t} - p N_{i,t} \frac{A_1 i + A_2}{t}.
\]
(7)

Recurrence relations for the expected values of \(N_{i,t}\) can be derived by taking the
expectation of the above equations. To solve these relations, we use the following lemma
on real sequences, which is Lemma 3.1 from [2].
Lemma 1. If \((\alpha_t), (\beta_t)\) and \((\gamma_t)\) are real sequences satisfying the relation

\[
\alpha_{t+1} = \left(1 - \frac{\beta_t}{t}\right) \alpha_t + \gamma_t,
\]

and \(\lim_{t \to \infty} \beta_t = \beta > 0\) and \(\lim_{t \to \infty} \gamma_t = \gamma\), then \(\lim_{t \to \infty} \frac{\alpha_t}{t}\) exists and equals \(\frac{\gamma}{1 + \beta}\).

Applying this lemma with \(\alpha_t = \mathbb{E}(N_{0,t})\), \(\beta_t = pA_2\), and \(\gamma_t = 1\) gives that \(\mathbb{E}(N_{0,t}) = c_0 t + o(t)\) with \(c_0\) as in (3). For \(i > 0\), the lemma can be inductively applied with \(\alpha_t = \mathbb{E}(N_{i,t})\), \(\beta_t = p(A_1 i + A_2)\), and \(\gamma_t = \mathbb{E}(N_{i-1,t})\frac{A_1(i-1) + A_2}{t}\) to show that \(\mathbb{E}(N_{i,t}) = c_i t + o(t)\), where

\[
c_i = c_{i-1} \frac{A_1(i-1) + A_2}{1 + p(A_1 i + A_2)}.
\]

It is easy to verify that the expression for \(c_i\) as defined in (3) and (4) satisfies this recurrence relation.

2.2 Concentration

We prove concentration for \(N_{i,t}\) when \(i \leq i_f\) by using a relaxation of Azuma-Hoeffding martingale techniques. The random variables \(N_{i,t}\) do not a priori satisfy the \(c\)-Lipschitz condition: it is possible that a new node may fall into many overlapping regions of influence. Nevertheless, we will prove that deviation from the \(c\)-Lipschitz condition occurs with exponentially small probability. The following lemma gives a bound for \(|N_{i,t+1} - N_{i,t}|\) which holds with extreme probability.

Lemma 2. We prove for all \(0 \leq t \leq n - 1\) the following inequalities hold.

i) \(|N_{i,t+1} - N_{i,t}| \leq 2(A_1 i + A_2) \log^2 n\), for \(0 \leq i \leq t\).

ii) \(|N_{i,t+1} - N_{i,t}| \leq 2(A_1 i + A_2)\), for \(\log^2 n < i \leq t\).

Proof. Fix \(t\), let \(i, j \leq t\), and let \(X_j(i, t)\) denote the indicator variable for the event that \(v_j\) has degree \(i\) at time \(t\) and \(v_{t+1}\) links to \(v_j\). Thus,

\[
N_{i,t+1} - N_{i,t} = \sum_{j=1}^{t} X_j(i-1, t) - \sum_{j=1}^{t} X_j(i, t),
\]

and so

\[
|N_{i,t+1} - N_{i,t}| \leq \max \left( \sum_{j=1}^{t} X_j(i-1, t), \sum_{j=1}^{t} X_j(i, t) \right). \tag{8}
\]

Let \(Z_j(i, t)\) denote the indicator variable for the event that \(v_{t+1}\) is chosen in the cap of area \((A_1 i + A_2)/t\) around node \(v_j\). Clearly, if \(X_j(i, t) = 1\), then \(Z_j(i, t) = 1\) as well,
so \( X_j(i, t) \leq Z_j(i, t) \). Thus, to bound \(|N_{i,t+1} - N_{i,t}|\) it suffices to bound the values of \( Z(i, t) \), where

\[
Z(i, t) = \sum_{j=1}^{t} Z_j(i, t).
\]

The variables \( Z_j(i, t) \) for \( j = 1, \ldots, t \) are pairwise independent. To see this, we can assume the position of \( v_{t+1} \) to be fixed. Then, the value of \( Z_j(i, t) \) depends only on the position of \( v_j \). Since the position of each node is chosen independently and uniformly, the value of \( Z_j(i, t) \) is independent from the value of any other \( Z_{j'}(i, t) \) where \( j \neq j' \). Therefore, \( Z(i, t) \) is the sum of independent Bernoulli variables with probability of success equal to

\[
P(Z_j(i, t) = 1) = \frac{A_1 i + A_2}{t}.
\]

Using Chernoff’s inequalities (see, for instance Theorem 2.1 [6]), we can show that \( Z(i, t) < A_1 i + A_2 + (A_1 i + A_2) \log^2 n < 2(A_1 i + A_2) \log^2 n \), and \( Z(i, t) < 2(A_1 i + A_2) \) if \( i > \log^2 n \). Using these bounds, the proof now follows since by (8),

\[
|N_{i,t+1} - N_{i,t}| \leq \max(Z(i - 1, t), Z(i, t)).
\]

\[\square\]

To sketch the technique of the proof of Theorem 1, we consider \( N_{0,t} \), the number of nodes of in-degree zero. We use the supermartingale method of Pittel et al. [9], as described in [10].

**Lemma 3.** Let \( G_0, G_1, \ldots, G_n \) be a random graph process and \( X_t \) a random variable determined by \( G_0, G_1, \ldots, G_t \), \( 0 \leq t \leq n \). Suppose that for some real \( \beta \) and constants \( \gamma_i \),

\[
E(X_t - X_{t-1}|G_0, G_1, \ldots, G_{t-1}) < \beta
\]

and

\[
|X_t - X_{t-1} - \beta| \leq \gamma_i
\]

for \( 1 \leq t \leq n \). Then for all \( \alpha > 0 \),

\[
P(\text{For some } t \text{ with } 0 \leq t \leq n : X_t - X_0 \geq t\beta + \alpha) \leq \exp\left(-\frac{\alpha^2}{2 \sum \gamma_j^2}\right).
\]

**Theorem 2.** We have for every \( t, 1 \leq t \leq n \)

\[
N_{0,t} = \frac{t}{1 + A_2 p} + O(n^{1/2} \log^3 n).
\]

**Proof.** We first transform \( N_{0,t} \) into something close to a martingale. Consider the following real-valued function

\[
H(x, y) = x^{pA_2} y - \frac{x^{1+pA_2}}{1 + pA_2}.
\]
(note that we expect $H(t, N_{0,t})$ to be close to zero). Let $w_t = (t, N_{0,t})$, and consider the sequence of random variables $(H(w_i) : 1 \leq i \leq n)$. The second-order partial derivatives of $H$ evaluated at $w_t$ are all $O(1)$. Therefore, we have

$$H(w_{t+1}) - H(w_t) = (w_{t+1} - w_t) \cdot \text{grad} \ H(w_t) + O(1),$$

where “$\cdot$” denotes the scalar product and \( \text{grad} \ H(w_t) = (H_x(w_t), H_y(w_t)) \).

Observe that, from our choice of $H$,

$$\mathbb{E}(w_{t+1} - w_t \mid G_t) \cdot \text{grad} \ H(w_t) = 0,$$

since $H$ was chosen so that $H(w)$ is constant along every trajectory $w$ of the differential equation that approximates the recurrence relation (6).

Hence, taking the expectation of (10) conditional on $G_t$, we obtain that

$$\mathbb{E}(H(w_{t+1}) - H(w_t) \mid G_t) = O(1).$$

From (10), noting that \( \text{grad} H(w_t) = (p A_2 t^{p A_2 - 1} N_{0,t} - t^{p A_2}, t^{p A_2}) \), and using Lemma 2 to bound the change in $N_{0,t}$, we have that \( \text{wep} \)

$$|H(w_{t+1}) - H(w_t)| \leq t^{p A_2} 2(A_1 i + A_2) \log^2 n + O(t) = O(t \log^2 n).$$

Now we may apply Lemma 3 to the sequence $(H(w_i) : 1 \leq i \leq n)$, and symmetrically to $(-H(w_i) : 1 \leq i \leq n)$, with $\alpha = n^{3/2} \log^3 n$, $\beta = O(1)$ and $\gamma_t = O(t \log^2 n)$, to obtain that \( \text{wep} \)

$$|H(w_t) - H(w_0)| = O(n^{3/2} \log^3 n).$$

As $H(w_0) = 0$, this implies from the definition (9) of the function $H$, that \( \text{wep} \)

$$N_{0,n} = \frac{n}{1 + p A_2} + O(n^{1/2} \log^3 n).$$

(11)

We may repeat the argument as in the proof of Theorem 2 for $N_{i,t}$ with $i \geq 1$. We omit the details here, which will follow in the long version of the paper.

### 2.3 In-degree of given node

PA model as analyzed in [3]. We may use Theorem 2.1 of [3] to describe the distribution of the in-degree of any given node. At any time-step $t$ In contrast to the large-scale behaviour of the degree distribution described in the previous subsection, here we focus on the distribution of the in-degree of an individual node. The indicator variable $Y_t$ for the increase in $d^-(v,t)$ by receiving a link from $v_{t+1}$ is Bernoulli $\text{Be}(p(A_1 d^-(v,t) + A_2)/t)$. Thus,

$$\mathbb{E}(d^-(v,t+1) \mid G_t) = d^-(v,t) + \frac{p (A_1 d^-(v,t) + A_2)}{t}.$$

(12)
This is very similar to the growth of the degree in the Preferential Attachment model as analyzed in [3]. As in the PA model, a "rich get richer" principle applies for the in-degrees, and the richer nodes are those that were born first. Theorem 2.1 of [3] can be used to obtain results on the concentration of $N_{i,t}$, but the methods employed in the previous sections give a stronger result.

The results on the distribution of $d^-(v,n)$ are summarized in parts (a) and (b) of the theorem below (use Theorem 2.2 of [3] with minor reworking). Part (c) will be discussed in the next section, and used to establish the concentration of the edges of $G_t$.

**Theorem 3.** Let $\omega = \log n$ and let $l^* = n^{\min\{PA_1,1/2\}}/\omega^4$. For $0 < PA_1 < 1$,

(a) For $\omega^8 \leq j \leq (n - n/\omega)$ and $0 \leq l \leq l^*$ or for $(n - n/\omega) < j < n$ and $l = 0, 1$,

$$\mathbb{P}(d^-(v_j, n) = l) = \left(1 + O(1/\omega^2)\right) \left(n/j\right)^{PA_1} \left(1 - \left(n/j\right)^{PA_1} \left(1 + O(1/\omega^2)\right)\right)^l.$$

(b) For $(n - n/\omega) < j < n$ and $l \geq 2$,

$$\mathbb{P}(d^-(v_j, n) = l) = O(l^{PA_1 - 1}/\omega^4).$$

(c) For all $K > 0$,

$$\mathbb{P}(\text{There exists } j \leq n: d^-(v_j, n) \geq K\omega^2(n/j)^{PA_1}) = O\left(n^{-K\omega^{-2}}\right).$$

Theorem 3(c) implies that $\text{aas}$ the maximum in-degree of node $v_j$ is at most $(n/j)^{PA_1}K\omega^2$. Conditional on this, (a) and (b) characterize the distribution of $d^-(v_j, n)$ for all $j \geq \omega^8$ when $PA_1 \leq 1/2$ and for $j \geq \omega^8 n^{PA_1 - 1/2}$ when $PA_1 > 1/2$.

### 3 The number of edges of $G_t$

We derive a concentration result for the number of edges in graphs generated by the SPA model. Let $M_t = |E_t|$, the number of edges in $G_t$, and let $m_t = \mathbb{E}(M_t)$. Then we have that

$$\mathbb{E}(M_{t+1} | M_t) = M_t + \sum_{j=1}^{t} p A_1 d^-(v_j, t) + A_2 \frac{t}{t} = M_t + pA_1 M_t + pA_2,$$

and so $m_1 = 0$, and for $t \geq 1$,

$$m_{t+1} = m_t \left(1 + \frac{PA_1}{t}\right) + pA_2.$$

The (first-order) solutions of this recurrence are

$$m_n \sim \begin{cases} \frac{pA_2}{1-PA_1} n, & \text{if } PA_1 < 1 \\ n \log n, & \text{if } PA_1 = 1. \end{cases}$$
Theorem 4. If \( pA_1 < 1 \), then the number of edges is concentrated around its expected value:
\[
M_n = m_n(1 + o(1)).
\]

The following lemma (whose proof is left to the long version of the paper) is used in the proof of Theorem 4, and proves Theorem 3 (c).

Lemma 4. For all \( v_j, j > 0 \) and \( K > 0 \),
\[
P(d^-(v_j, n) \geq K \log^2 n(n/j)^{pA_1}) = O(n^{-K\epsilon^{-18}}).
\]

Proof of Theorem 4. We count the number of edges by counting the in-degree of nodes. Our approach is as follows: by Theorem 1 \( \mu(n) = \sum_{i \leq i_f} \mathbb{E}N_i(n) \) and let \( A \) be the solution of \( (n/A)^{pA_1} = i_f \) and let \( \omega' = (K \log^2 n)^{1/(pA_1)} \) be the solution of
\[
\left( \frac{t}{\omega'} \right)^{pA_1} K \log^2(n) = \left( \frac{n}{A} \right)^{pA_1},
\]
where \( K \geq 4e^{18} \). From Lemma 4, with probability \( 1 - O(n^{-3}) \) no node \( v \geq a\omega' \) has degree exceeding \( i_f \). Let \( \mu(n) = \sum_{i \leq i_f} \mathbb{E}N_i(n) \), and let \( \lambda(n) = \sum_{j=1}^{a\omega'} d^-(v_j, n) \). We prove, conditional on Lemma 4, that \( \lambda(n) = o(m_n) \) and thus the number of edges is concentrated around \( m_n \). We have that for \( pA_1 < 1 \)
\[
\lambda(n) = \sum_{j=1}^{a\omega'} d^-(v_j, n)
\leq K \omega'^2 \sum_{j=1}^{a\omega'} \left( \frac{n}{j} \right)^{pA_1}
= O(1/(1 - pA_1)) \log^{2/(pA_1)}(n)n^{pA_1}a^{1-pA_1}
= O(1/(1 - pA_1)) \log^{2/(pA_1)}(n) + 4(1 - pA_1)/(6pA_1 + 2)n_6^{pA_1+1}/6^{pA_1+2}
= o(n).
\]

However, \( \mu(t) \geq ct \) for some constant \( c > 0 \).

4 A geometric small world property

In Section 2 it was shown that the number of nodes in a graph generated by the SPA model of in-degree zero in \( G_n \) is linear in \( n \). Also, with positive probability a new node will land in an area of \( S \) not covered by any influence regions, and thus have out-degree zero. Therefore, the underlying undirected graph of \( G_n \) is not connected. In fact, we
expect that for the majority of distinct pairs \( u, v \), there will not be a directed path from \( u \) to \( v \). Since this is a property also observed in the web graph, it does not detract from the SPA model, but rather indicates that we should consider another variable rather than diameter to indicate a “small world” property. Thus, we focus on the (geometric) distance, in \( S \), spanned by the links.

For a pair of points \( u, v \in S \), let \( L(u, v) \) be the length of the shortest curve embedded in the surface of \( S \) that connects \( u \) and \( v \). Define

\[
L_t = \sum_{(v_t, v_i) \in E_t} L(v_t, v_i);
\]

that is, \( L_t \) is the sum of the lengths of new edges added at time \( t \) in the SPA model. Note that \( L_t \) is a continuous random variable.

**Theorem 5.** For the expectation of \( L_t \),

\[
\mathbb{E}(L_t) = \Theta \left( t^{(1-pA_1)} \right).
\]

To prove Theorem 5 we need the following lemma whose (straightforward) proof is omitted.

**Lemma 5.** Let \( u \) be chosenuar from a cap with centre \( v \) and area \( \alpha \). If \( X \) is the distance between \( u \) and \( v \), measured over the surface of \( S \), then

\[
\mathbb{E}(X) = \frac{2\sqrt{\alpha \pi}}{3}.
\]

**Proof of Theorem 5** Define

\[
Z_{j,t} = \begin{cases} L(v_t, v_j) & \text{if } (v_t, v_j) \in E_t \\ 0 & \text{else}. \end{cases}
\]

Then \( L_t = \sum_{j=1}^{t-1} Z_{j,t} \). Let \( B_{t,j} \) be the event that \((v_t, v_j) \in E_t \). Then using Lemma 5 we have that

\[
\mathbb{E}(Z_{j,t+1} \mid G_t) = \mathbb{P}(B_{t,j})\mathbb{E}(Z_{j,t+1} \mid G_t, B_{t,j}) + \mathbb{P}(\overline{B}_{t,j})\mathbb{E}(Z_{j,t+1} \mid G_t, \overline{B}_{t,j})
= \mathbb{P}(B_{t,j})\mathbb{E}(L((v_{t+1}, v_j) \mid G_t)
= \left( p A_1 d^-(v_j, t) + A_2 \right) \left( \frac{2}{3} \sqrt{\frac{A_1 d^-(v_j, t) + A_2}{\pi t}} \right)
= \frac{2p}{3\sqrt{\pi}} \left( \frac{A_1 d^-(v_j, t) + A_2}{t} \right)^{3/2},
\]

where the second last equality follows by Lemma 5 and the definition of the model, and the second equality follows from the definition of \( Z_{j,t+1} \). Thus

\[
\mathbb{E}(L_{t+1} \mid G_t) = \sum_{k=0}^{t} \sum_{\{j:d^-(v_j, t) = k\}} \mathbb{E}(Z_{j,t+1}|G_t) = \frac{2p}{3\sqrt{\pi}} \sum_{k=0}^{t} \left( \frac{A_1 k + A_2}{t} \right)^{3/2} N_{k,t}.
\]
Taking expectations on both sides, and using that $c_k = ck^{-(1+pA_1)}(1+o(1))$, we have that

\[ \mathbb{E}(L_{t+1}) = \frac{2p}{3\sqrt{\pi}} \sum_{k=0}^{t} \left( \frac{A_1 k + A_2}{t} \right)^{3/2} \mathbb{E}(N_{k,t}) \]

\[ = \frac{2p}{3\sqrt{\pi t}} \sum_{k=0}^{t} (A_1 k + A_2)^{3/2} c_k \]

\[ = \frac{2pc}{3\sqrt{\pi t}} \int_0^t x^{1/2-1/(pA_1)}(1 + o(1))dx \]

\[ = \Theta(t^{1-1/(pA_1)}), \]

where the second equality follows by Theorem 1. The last step is justified since it can be shown that the $o(1)$ term in the integrand is in fact $O(x^{-\epsilon})$ for some $\epsilon > 0$. \hfill \Box

Theorem 5 contrasts with the analogous result for graphs generated with a similar process to the SPA model, but where all influence regions have area $d/t$ for $d > 0$ a constant. We call this a threshold model. In the threshold model, $\mathbb{E}(L_t)$ decreases much faster than for the SPA model with $p$ large, such as when $p > 2/3$ and $A_1 = 1$. For example, if $pA_1 = 1$, then $\mathbb{E}(L_t) = O(1)$.

**Theorem 6.** In the threshold model with areas of influence $d/t$, where $d$ is a constant,

\[ \mathbb{E}(L_t) \sim ct^{-1/2}. \]

**Proof.** With the same notation as in the proof of Theorem 5 and using Lemma 5, we have that

\[ \mathbb{E}(Z_{j,t+1} \mid G_t) = \mathbb{P}(B_{j,t+1}) \mathbb{E}(L(v_{t+1}, v_j) \mid B_{j,t+1}) \]

\[ = \frac{2d}{3t} \sqrt{\frac{d}{\pi t}}. \]

Hence,

\[ \mathbb{E}(L_{t+1} \mid G_t) = \sum_{i=1}^{t} \mathbb{E}(Z_{j,t+1} \mid G_t) = \frac{2d}{3} \sqrt{\frac{d}{\pi t}} = \Theta(t^{-1/2}). \]

Taking expectations completes the proof. \hfill \Box

## 5 Conclusions and further work

We have proved that graphs produced by the SPA model have some of the graph properties observed in real-world complex networks: a power law in-degree distribution, and constant average degree. In future work, we will investigate additional graph properties,
such as the expected length of a directed path between two nodes (when such a path exists), expansion properties, and spectral values. We are also interested in aspects suggesting self-similarity: is it true that the subgraph induced by all nodes that fall in a certain compact region of the sphere $S$ share some of the graph properties of the whole graph?

Several generalizations of this model may be proposed. An undirected version could be developed, where the link probability depends on the influence regions of both endpoints. In a more realistic model, both the addition of edges without adding a node and the deletion of edges and nodes should be incorporated. The effect of replacing $S$ with other underlying geometric spaces, either with boundaries or of higher dimension, would be interesting to investigate.

Last but not least, a realistic spatial model gives the possibility for reverse engineering of real-life networks: given a real-life network and assuming a spatial graph model by which the network was generated, it should be possible to give reliable estimates about the positions of the nodes in space. This direction has important applications to web graph clustering and development of link-based similarity measures.

References