

GEOMETRIC GRAPH PROPERTIES OF THE SPATIAL PREFERRED ATTACHMENT MODEL

JEANNETTE JANSSEN, PAWEŁ PRAŁAT, AND RORY WILSON

ABSTRACT. The spatial preferred attachment (SPA) model is a model for networked information spaces such as domains of the World Wide Web, citation graphs, and on-line social networks. It uses a metric space to model the hidden attributes of the vertices. Thus, vertices are elements of a metric space, and link formation depends on the metric distance between vertices. We show, 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. Precisely, the estimate is based on the number of common neighbours of a pair of vertices, a measure known as *co-citation*. To be able to calculate this estimate, we derive a precise relation between the number of common neighbours and metric distance. We also analyze the distribution of *edge lengths*, where the length of an edge is the metric distance between its end points. We show that this distribution has three different regimes, and that the tail of this distribution follows a power law.

1. INTRODUCTION

Thanks to the World Wide Web and its hyperlinked structure, more and more information is becoming available in the form of a networked information space: a collection of information entities (documents, scientific papers, Web pages, individuals in a social network), connected by links between pairs of entities (references, citations, hyperlinks, “friend” relationships). Studies of various networked information spaces have given convincing evidence that a significant amount of information about the entities represented by the vertex can be derived from the graph representing the link structure. This has led to the application of graph-theoretical techniques to such graphs, with the aim of developing methods to understand the link structure and mine its information.

An important step in understanding the link structure is the development of a graph model; a stochastic process that models the link formation. The first generation of graph models was mainly aimed at explaining the graph-theoretical properties observed in real-life networks. In such models, vertices are considered anonymous, and link formation is only influenced by the current link structure. An example is the seminal model by Barabási and Albert in [2] based on the principle of *preferential attachment*: each new vertex attaches randomly to a prescribed number of existing vertices, with a link probability proportional to the degree, so vertices of high degree are more likely to receive a link from the new vertex.

Key words and phrases. Node similarity, co-citation, bibliographic coupling, link analysis, complex networks, spatial graph model, SPA model.

The authors gratefully acknowledge support from NSERC and MITACS grants.

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.

In this paper, we focus on the Spatial Preferred Attachment (SPA) model, proposed in [1], and analyze the relationship between the link structure of graphs produced by the model, and the relative positions of the vertices in the metric space. 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. A model for on-line social networks based on similar principles can be found in [4, 5].

A number of spatial models have been proposed recently [6, 10, 11, 19, 12]. In these models, as in the SPA model, the relationship between spatial distance and link formation is determined by a threshold function: a link is possible if vertices are within a prescribed threshold distance of each other, and impossible otherwise. However, for these models the threshold distance remains constant throughout the process, and does not depend on the degree, and decrease with time, as in the SPA model.

A different class of graphs explores the interplay between distance and edge likelihood—with associated graph properties—with more involved mechanisms than simple thresholds: for example, in [23], each new vertex is born with m edges, each joining a neighbour with probability proportional to the in-degrees and a function of the distance between them. Variations include the deterministic model [21] in which edges are formed based on the “utility” for the nodes in question, utility incorporating both in-degree and distance; in [9], the model demands that the number of nodes per unit volume is constant, and an analysis on the distribution of edge lengths is also included. Beyond the creation of models, [17] takes a closer look at the concept of complex networks having an underlying geometry. For a recent survey of spatial models, see [14].

Our first main result shows that, for the SPA model, the number of common in-neighbours between a pair of vertices can, in many cases, be used to estimate the distance between the vertices. Since the metric distance is assumed to represent the similarity or “closeness” of the entities represented by the vertices, this means that it is possible to estimate similarity between vertices by looking at the graph only, i.e. without considering the underlying reality represented by the metric

space. The number of common in-neighbours in a citation graph is known in library science as the measure of *co-citation*, and is one the earliest measures of graph-based similarity, proposed by Small in 1973 in [22]. Co-citation, and the related measure of bibliographic coupling (from [15]) based on the number of common out-neighbours, are widely used link similarity measures for scientific papers, via the citation graph, for Web pages, and others [3, 8, 20, 18].

The question of determining similarity between vertices is one that is central to many link mining applications. It is an important tool in searching, by finding documents or Web pages that are similar to a given target document. It can also be used as the basis to identify *communities*, or clusters, of similar vertices. A purely graph-based measure of similarity can be used as a complementary indication of similarity between vertices when other information is unreliable (as is often the case in the World Wide Web), largely unavailable (as in some biological networks and online social networks), or protected by privacy laws (as in networks representing phone calls or bank transactions).

Our result on the relationship between number of common neighbours and metric distance is derived theoretically through an analysis of the SPA model. The analytic result is asymptotic in the size of the graph. In order to test the result on realistic graph sizes, we performed simulations for graphs of 100,000 vertices, with various parameter choices. The simulations show that the real distance and the predicted distance from the number of common neighbours are in very good agreement.

Our second main result determines the distribution of the *edge lengths*, where the length of an edge is the metric distance between its end points. Edge length is a metric property of a graph feature, and edge length distribution is a combined metric/graph property which is unique to spatial graph models. In the SPA model, the maximum length of an edge is determined by the size of the sphere of influence of its destination vertex, and this size is determined by the degree of the vertex. Since the degrees follow a power law, we might expect that the edge length distribution follows a power law. We show, both through theoretical results and simulations, that the situation is slightly more complex. In fact, we present clear evidence that, for a certain combination of model parameters, there are three different regimes of the distribution. For the smallest edge lengths, the cumulative edge length distribution is constant: almost all edges fall in this category. In the mid range, we have a power law with coefficient between 0 and 1, and in the tail, we have a power law with exponent greater than 1.

In Section 2, we describe the SPA model and derive some properties on the degree of a vertex which we will need to establish our results. In Section 3, we give the result on common in-neighbours and metric distance, and present the simulation results. In Section 4, we state our theorem on edge length distribution, and present the edge length distribution as obtained through simulations for various parameters. In Section 5, we give proofs of all the main theorems.

2. THE SPA MODEL

We start by giving a precise description of the SPA model, and deriving some facts about the degrees of the vertices, which we will need to prove our main results. In [1], 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 ,

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

The torus metric thus “wraps around” the boundaries of the unit square; this metric was chosen to eliminate boundary effects. Let c_m be the constant of proportionality of volume used with the 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 . 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:

$$|S(v, t)| = \frac{A_1 \deg^-(v, t) + A_2}{t}, \quad (1)$$

or $S(v, t) = S$ and $|S(v, t)| = 1$ if the right-hand-side of (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 note that, to avoid the resulting graph becoming too dense, the parameters must be chosen so that $pA_1 < 1$, as explained in [1]. In this paper, we assume that the parameters meet this condition. Also, the original model as presented in [1] 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 .

We now introduce some more definitions. In the rest of the paper, unless otherwise stated we will assume all asymptotics to refer to n going to infinity, where n is the end time of the growth process, and thus the final size of the graph. (As explained above, Theorem 2.1 is an exception.) We say that an event holds *asymptotically almost surely* (a.a.s.) if the probability that it holds tends to one as n goes to infinity. Similarly, we will use *with extreme probability* (w.e.p.) if the event holds with probability at least $1 - \exp(-\Theta(\log^2 n))$. Thus, if we consider a polynomial number of events that each holds w.e.p., then w.e.p. all events hold.

It was shown in [1] that the SPA model produces graphs with a power law degree distribution, with exponent $1 + 1/(pA_1)$. In [7], the (directed) diameter

of the model was investigated. For the results of this paper, we need a precise expression for the expected in-degree of each vertex.

Theorem 2.1. *Let $\omega = \omega(t)$ be any function tending to infinity together with t . The expected in-degree at time t of a vertex v_i born at time $i \geq \omega$ is given by*

$$\mathbb{E}(\deg^-(v_i, t)) = (1 + o(1)) \frac{A_2}{A_1} \left(\frac{t}{i}\right)^{pA_1} - \frac{A_2}{A_1}. \quad (2)$$

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

$$X(v_i, t) = \deg^-(v_i, t) + \frac{A_2}{A_1}. \quad (3)$$

It follows immediately from the definition of the process that

$$X(v_i, t+1) = \begin{cases} X(v_i, t) + 1, & \text{with probability } \frac{pA_1 X(v_i, t)}{t} \\ X(v_i, t), & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} \mathbb{E}(X(v_i, t+1) \mid X(v_i, t)) &= (X(v_i, t) + 1) \frac{pA_1 X(v_i, t)}{t} + X(v_i, t) \left(1 - \frac{pA_1 X(v_i, t)}{t}\right) \\ &= X(v_i, t) \left(1 + \frac{pA_1}{t}\right), \end{aligned}$$

and so

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

Since all vertices start with in-degree zero, $X(v_i, i) = \frac{A_2}{A_1}$. Since $i \geq \omega$, one can use this to get

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

and the assertion follows from (3). \square

Theorem 2.1 states that the *expected* in-degree of an individual vertex born at time i is asymptotically equal to $\frac{A_2}{A_1} \left(\frac{t}{i}\right)^{pA_1} - \frac{A_2}{A_1}$, with an error term of order $o((t/i)^{pA_1})$. (The asymptotics assume that t is going to infinity, and i is a growing function of t .) However, the in-degree of an individual vertex is not concentrated around its expected value. This is due to variation happening shortly after birth; whether or not the vertex receives in-links in the first few time steps after its

birth greatly affects the size of its sphere of influence throughout the process, and therefore its final in-degree.

We can circumvent this difficulty by considering the final in-degree of the vertex, and infer the growth history of the in-degree from there. Namely, from the in-degree of the vertex at end time n , we can obtain sharp bounds on the in-degree of the vertex during most of the process. This is expressed in the following theorem. First, define an injective function $f : \mathbb{R} \rightarrow \mathbb{R}$ by

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

so $f(i)$ is the expected in-degree, at time n , of a vertex born at time i (up to a multiplicative factor of $(1 + o(1))$). Thus $f^{-1}(k)$ is the birth time of a vertex of final in-degree k , had the in-degree of the vertex remained close to its expected value during its entire lifetime. Moreover, the (asymptotic) expected in-degree at time t of a vertex born at time i can be given as $(A_2/A_1)f(i)/f(t)$ (provided that $i = i(n)$ tends to infinity with n). Thus, if a vertex of final in-degree k has in-degree growth close to its expected value, then

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

will be the approximate time when that vertex has in-degree a . The precise statement and proof of this discussion follows below in Theorem 2.2, the main result of this section.

Theorem 2.2. *Let $\omega = \omega(n)$ be any function tending to infinity together with n . The following statement holds a.a.s. for every vertex v for which $\deg^-(v, n) = k = k(n) \geq \omega \log n$. Let $i = f^{-1}(k)$, and let*

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

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

$$\deg^-(v, t) = (1+o(1)) \frac{A_2}{A_1} \left(\frac{t}{i} \right)^{pA_1} = (1+o(1)) \frac{A_2}{A_1} \cdot \frac{k}{f(t)} = (1+o(1)) k \left(\frac{t}{n} \right)^{pA_1}. \quad (4)$$

The theorem implies that once a given vertex accumulates $\omega \log n$ in-neighbours, the rest of the process (until time-step n) can be predicted with high probability; in fact, a.a.s. we get a concentration around the expected value. Let us mention that it seems that the ω factor is needed to get a concentration result. However, without this factor, the order of the in-degrees still can be predicted: once the vertex has $\log n$ in-neighbours, we can bound the in-degree of this vertex so that the ratio between upper and lower bounds would a.a.s. be a constant.

In order to prove Theorem 2.2, we need strong results on the concentration of the in-degree throughout the process. These results, and the proof of the theorem, are given in Section 5.

3. NUMBER OF COMMON NEIGHBOURS AND SPATIAL DISTANCE

The principles of the SPA model make it plausible that vertices that are close together in space will have a relatively high number of common neighbours. Namely, if two vertices are close together, their spheres of influence will overlap during most of the process, and any new vertex falling in the intersection of both spheres has the potential to become a common neighbour. Thus, the number of common neighbours (co-citation) should lead to a reliable measure of closeness in the metric space. In this section, we will quantify the relation between spatial distance and number of common in-neighbours, and show how it can be used to estimate distance.

The term “common neighbour” here refers to common in-neighbours. Precisely, a vertex w is a common neighbour of vertices u and v if there exist directed links from w to u and from w to v . Note that in our model this can only occur if w is younger than u and v , and, at its birth, w lies in the intersection of the spheres of influence of u and v . We use $cn(u, v, t)$ to denote the number of common in-neighbours of u and v at time t .

Theorem 3.1 distinguishes three cases. The division into cases is based on the trend, as shown in Theorem 2.2, that spheres of influence tend to shrink over time. Thus, once the spheres of influence of two vertices have become disjoint, and their boundaries have some distance between them, it is not likely that they will overlap at any time after that. The cases therefore are distinguished by how the spheres of influence of u and v overlap, and when or whether they become disjoint. Figure 1 gives a pictorial representation of the three cases. Consider two vertices u and v so that v has smaller in-degree at time n than u . Thus, the sphere of influence of v tends to be smaller than that of u , and the likely birth time of u is before that of v .

In Case 1, u and v are so far apart that their spheres of influence never overlap, except maybe for a negligible initial time period near their birth. In this case, no vertex can fall in the spheres of influence of both u and v , and thus u and v will acquire no common neighbours after the initial time period. Thus, they will have negligibly few common neighbours. In this case again, accurate prediction of the spatial distance between u and v is not possible: if u and v have very few common neighbours, we can only give a lower bound on their distance.

In Case 2, u and v are so close that the sphere of influence of v is contained within the sphere of influence of u for almost all of its existence. In this case, the number of common neighbours of u and v is a constant proportion of the degree of v , due to the fact that each new vertex linking to v will automatically be within the sphere of influence of u , and thus can link to u as well (and does so with probability p .) This means that u and v are too close for accurate prediction: if $cn(u, v, n)$ and $\deg^-(v, n)$ differ by a factor close to p we can only give an upper bound on the spatial distance between u and v .

In Case 3, the sphere of influence of v is contained in that of u near the birth of v , but the spheres become disjoint before the end of the process. The moment at which the separation occurs can be determined fairly precisely, and depends heavily on the distance between u and v . Thus, for this case we have a formula

Case	Near birth of v	Near end of process	
1			Too far
2			Too close
3			Just right

FIGURE 1. The three cases of Theorem 3.1

for the number of common neighbours which involves the distance between u and v , and the in-degree of both u and v at the end of the process. Reversing the formula, we can obtain a reliable estimate for the distance between u and v from the observable graph properties $cn(u, v, n)$, $\deg^-(u)$ and $\deg^-(v)$.

Theorem 3.1. *Let $\omega = \omega(n)$ be any function tending to infinity together with n . The following holds a.a.s. Let v_k and v_ℓ be vertices such that*

$$k = \deg(v_k, n) \geq \deg(v_\ell, n) = \ell \geq \omega^2 \log n$$

in a graph generated by the SPA model. Let $d = d(v_k, v_\ell)$ be the distance between v_k and v_ℓ in the metric space. Finally, let $T = f^{-1}(\ell/(\omega \log n))$. Then,

Case 1. *If $d \geq \varepsilon(\omega \log n/T)^{1/m}$ for some $\varepsilon > 0$, then*

$$cn(v_\ell, v_k, n) = O(\omega \log n).$$

Case 2. *If $k \geq (1 + \varepsilon)\ell$ for some $\varepsilon > 0$ and*

$$d \leq \left(\frac{A_1 k + A_2}{c_m n} \right)^{1/m} - \left(\frac{A_1 \ell + A_2}{c_m n} \right)^{1/m} = \Theta \left(\left(\frac{k}{n} \right)^{1/m} \right), \quad (5)$$

then

$$cn(v_\ell, v_k, n) = (1 + o(1))p\ell.$$

If $k = (1 + o(1))\ell$ and $d \ll (k/n)^{1/m} = (1 + o(1))(\ell/n)^{1/m}$, then $cn(v_\ell, v_k, n) = (1 + o(1))p\ell$ as well.

Case 3. If $k \geq (1 + \varepsilon)\ell$ for some $\varepsilon > 0$ and

$$\left(\frac{A_1 k + A_2}{c_m n}\right)^{1/m} - \left(\frac{A_1 \ell + A_2}{c_m n}\right)^{1/m} < d \ll (\omega \log n/T)^{1/m}, \quad (6)$$

then

$$cn(v_\ell, v_k, n) = C i_k^{-\frac{(pA_1)^2}{1-pA_1}} i_\ell^{-pA_1} d^{-\frac{mpA_1}{1-pA_1}} \left(1 + O\left(\left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right)\right), \quad (7)$$

where $i_k = f^{-1}(k)$ and $i_\ell = f^{-1}(\ell)$ and $C = pA_1^{-1} A_2^{\frac{1}{1-pA_1}} c_m^{-\frac{pA_1}{1-pA_1}}$.

If $k = (1 + o(1))\ell$ and $\varepsilon(k/n)^{1/m} < d \ll (\omega \log n/T)^{1/m}$ for some $\varepsilon > 0$, then

$$cn(v_\ell, v_k, n) = \Theta\left(i_k^{-\frac{(pA_1)^2}{1-pA_1}} i_\ell^{-pA_1} d^{-\frac{mpA_1}{1-pA_1}}\right).$$

The importance of the theorem is that (7) gives a relationship between the distance between the vertices, their number of common neighbours, and their degrees. Since the number of common neighbours and the degrees are observable from the graph, the equation allows us to obtain an estimate for the (spatial) distance between the vertices using only basic graph parameters.

We tested the predictive power of our theoretical results on data obtained from simulations. The data was obtained from a graph with 100,000 vertices. The graph was generated from points randomly distributed in the unit square in \mathbb{R}^2 according to the SPA model described in Section 2, with $n = 100,000$ and $p = 0.95$, and $A_1 = A_2 = 1$.

First of all, we show that a blind approach to using the co-citation measure (number of common neighbours) does not work. In Figure 2 we plot spatial distance versus number of common neighbours without further processing. No relation between the two is apparent.

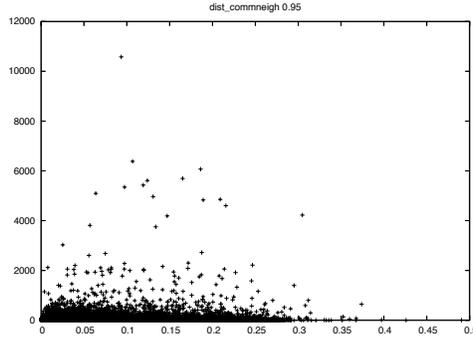


FIGURE 2. Actual distance vs. number of common neighbours.

Next, we apply Theorem 3.1 to estimate the spatial distance between two vertices, based on the number of common neighbours of the pair. (The spatial distance is actual distance between the point in the metric space, which for our simulation is the distance obtained from the Euclidean torus metric on the unit

square.) From Cases 1 and 2, we can only obtain a lower and upper bound on the distance, respectively. In order to eliminate Case 1 (too far), we consider only pairs that have at least 20 common neighbours. This reduces the data to 19,200 pairs. For pairs of vertices in Case 2 (too close), the number of common neighbours equals p times the lowest degree of the pair. In order to eliminate this case, we require that the number of common neighbours should be less than $p/2$ times the lowest degree of the pair. This reduces the data set to 2,400 pairs. We expect these pairs mainly to be in Case 3.

For pairs in Case 3, we can derive an estimate of the distance. Consider two such vertices v_ℓ and v_k , with final in-degree ℓ and k , respectively. We base our estimate on Equation 7, where we ignore the multiplicative $(1 + O((\frac{i_k}{i_\ell})^{pA_1/m}))$ error term. Namely, when k and ℓ are of the same order, then this expression is the average of the lower and upper bound as derived in the proof of the theorem, and when $\ell \ll k$ the term is asymptotically negligible. The estimated distance \hat{d} between nodes v_ℓ and v_k , given that their number of common neighbours equals N , is then given by

$$\hat{d} = C' i_k^{-\frac{pA_1}{m}} i_\ell^{-\frac{1-pA_1}{m}} N^{-\frac{1-pA_1}{mpA_1}},$$

where $i_k = f^{-1}(k)$ and $i_\ell = f^{-1}(\ell)$ and $C' = (p/A_1)^{\frac{1-pA_1}{mpA_1}} A_2^{\frac{1}{mpA_1}} c_m^{-\frac{1}{m}}$.

Figure 3 shows actual vs. estimated distance for these pairs. The estimated distance (on the y -axis), is computed using only data obtainable from the graph: the in-degrees of both vertices, and their number of common neighbours. This is compared to the actual distance (on the x -axis), known from the simulation. We see almost perfect agreement between estimate and reality.

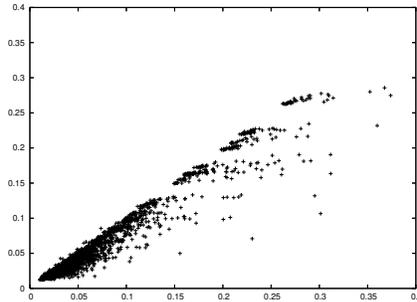


FIGURE 3. Actual distance (x -axis) vs. estimated distance (y -axis) for eligible pairs from simulated data, calculated using the in-degree of both vertices.

The figure shows that the scatter away from the diagonal is confined to points below the diagonal. This means that, for the corresponding pairs, the estimate \hat{d} is lower than the actual distance. This is due to the choice to base our estimate on the average between the lower bound obtained from t^- , the estimated time when the sphere of influence of v_ℓ first touches the boundary of the sphere of influence of v_k , and the upper bound derived from t^+ , when the spheres of influence of v_ℓ and v_k first become disjoint.

The probability that a neighbour of v_ℓ born between t^- and t^+ becomes a common neighbour of v_k and v_ℓ depends on the fraction of the sphere of influence of v_ℓ which lies inside the sphere of influence of v_k . If the curvature of the sphere of influence of v_k is negligible so that the boundary locally resembles a line, and if the sphere of influence of v_ℓ remains constant in size from t^- to t^+ , then the average is a good estimate. However, both assumptions are notably false: the curvatures of the spheres of influence of v_ℓ and v_k may well be of the same order, and the spheres of influence both shrink during the process. This implies that the fraction of the sphere of influence of v_ℓ inside the sphere of influence of v_k is smaller than assumed near time t^+ , and larger than assumed near t^- . Thus, the true expected number of common neighbours will likely be larger than indicated by the average. This leads to an underestimate of the distance (more common neighbours is interpreted as closer distance).

In order to test our interpretation of the error in the estimation, we based the estimator \hat{d} on a convex combination of the lower bound L on the numbers of common neighbours of vertices v_k and v_ℓ given by $L = p \deg^-(v_\ell, t^-)$ and the upper bound $U = p \deg(v_\ell, t^+)$. So the expected number of common neighbours is assumed to be $(1 - c)L + cU$, which gives an expression involving d . Solving for d gives our estimator \hat{d} . We found that the best value of c occurred when $c = 0.005$, which means that the lower bound based on time t^- gives the best indication of the true number of common neighbours.

The results for this adjusted estimator are given in Figure 4. As we can see, the estimator is still not perfect; we conjecture that this is because the value of c that gives the best estimate is not uniform over all pairs, but depends on the relative sizes of the spheres of influence of the pair in the critical time interval.

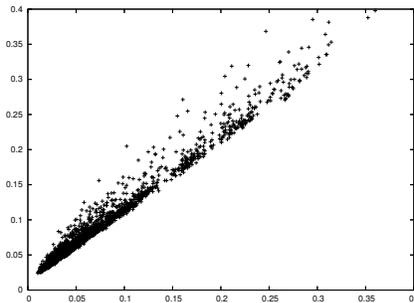


FIGURE 4. Actual distance (x -axis) vs. estimated distance (y -axis) for eligible pairs from simulated data, using the adjusted estimator.

4. EDGE LENGTH DISTRIBUTION

In this section we derive the edge length distribution; that is, the number of edges whose length is at least a given value x . The length of an edge is the (metric) distance between its endpoints. The edge distribution is a characteristic of spatial models. It will influence a number of graph properties, especially the diameter and the expansion properties. Long edges, even if they are rare, give the opportunity to jump to another locality in the metric space. It has been shown before (see,

for example, [16]) that a small number of long edges can reduce the average path length between vertices by a large factor.

In the SPA model, the degree distribution follows a power law, and the volume of the spheres of influence is proportional to the degree of a vertex. The radius of the sphere of influence determines the limit of the length of an edge to that vertex. Thus, we expect the edge lengths to follow a power law as well. These considerations lead us to consider all edges whose length exceeds a given value

$$r_\alpha = \left(\frac{n^{-\alpha}}{c_m} \right)^{1/m}.$$

(Recall that c_m is the volume of an m -dimensional ball of unit radius.) Namely, in this case we can limit our focus to those vertices whose sphere of influence has volume at least $n^{-\alpha}$.

Fix $\alpha > 0$. An edge $(v, w) \in E(G)$ will be called a *long edge* if the edge length $d(v, w) \geq r_\alpha$. We will study the random variable $e(\alpha)$, the number of long edges in the graph. Formally,

$$e(\alpha) = \left| \left\{ (v, w) \in E : d(v, w) \geq \left(\frac{n^{-\alpha}}{c_m} \right)^{1/m} \right\} \right|.$$

Theorem 4.1. *In the SPA Model with $1/2 < pA_1 < 1$, a.a.s. the number of long edges is given by*

$$e(\alpha) = \begin{cases} (1 + o(1)) \frac{pA_2}{1-pA_1} n, & \text{if } \alpha > 1 \\ (1 + o(1)) C n^{2 - \frac{1}{pA_1} + \alpha \frac{1-pA_1}{pA_1}}, & \text{if } 1 - \frac{pA_1}{4pA_1+2} < \alpha < 1, \end{cases} \quad (8)$$

where

$$C = \frac{\Gamma\left(\frac{A_2}{A_1} + \frac{1}{pA_1}\right) A_1^{\frac{pA_1}{1-pA_1}}}{\Gamma\left(\frac{A_2}{A_1}\right) (1-pA_1)} \left(\frac{(1-pA_1)^3}{2pA_1-1} A_1^{\frac{1-2pA_1}{(1-pA_1)pA_1}} + 1 - (pA_1)(1-pA_1) \right).$$

By [1], the total number of edges in graphs generated by the SPA model equals $(1 + o(1)) \frac{pA_2}{1-pA_1} n$. Thus, the first case of the theorem states that for $\alpha > 1$, $e(\alpha)$ is approximately equal to the total number of edges. To see why this is so, consider that, as α increases, the threshold for an edge to be classified as “long”, namely r_α , decreases. If $\alpha > 1$, then r_α is so small that almost all edges are long.

The next range for α , $1 - \frac{pA_1}{4pA_1+2} < \alpha < 1$, shows a linear relationship between $\log e(\alpha)$ and $\log r_\alpha$. Namely, $m \log r_\alpha = (1 + o(1))(-\alpha) \log n$, and thus for this range,

$$\begin{aligned} \log e(\alpha) &= (1 + o(1)) \left(2 - \frac{1}{pA_1} + \alpha \frac{1-pA_1}{pA_1} \right) \log n \\ &= (1 + o(1)) \left(\left(2 - \frac{1}{pA_1} \right) \log n - m \frac{1-pA_1}{pA_1} \log r_\alpha \right). \end{aligned} \quad (9)$$

Since $1/2 < pA_1 < 1$, the slope of the line giving the relationship between α and $\log e(\alpha)$ lies between 0 and 1.

The theorem does not include a claim about the tail of the edge distribution, when α becomes small, and thus r_α becomes relatively large. When $1 - pA_1 < \alpha \leq 1 - \frac{pA_1}{4pA_1+2}$, the main contribution to $e(\alpha)$ comes from vertices that have very high final degree (not moderately high, as before) and the long edges are created till the very end of the process. Unfortunately, the number of vertices of very high degree cannot be precisely controlled ; from [1] we only have upper bounds and lower bounds on the maximum degree that hold *w.e.p.* which differ by a factor of $\log^4 n$. Therefore, it seems unlikely that $e(\alpha)$ is concentrated in this case.

When $\alpha < 1 - pA_1$, *a.a.s.* long edges cannot be created at the end of the process but only until time $s = n^{\alpha/(1-pA_1)+o(1)}$. The main contribution to the number of long edges comes from those vertices that have very high degree at time s (and have very high final degrees, of course). By a similar argument as given above, the number of such vertices, and thus the value of $e(\alpha)$, is not likely to be highly concentrated.

A different problem occurs when $pA_1 < 1/2$. The main contribution in this case comes from vertices born at time $\Theta(n^\alpha)$ and the long edges must have been created when these vertices were still young, and had small degrees. Unfortunately, the behaviour of the random variable representing the degree of a vertex is not concentrated until the degree is $\omega \log n$. We expect $\Theta(n^\alpha)$ such edges but we cannot control the behaviour of these vertices until the degree is large enough.

The following theorem fills in the missing case when α is small. However, the results only apply to the *expected* value of $e(\alpha)$, and they give broad results about the order of the exponent, instead of the finer results of the previous theorem. The proofs of both theorems can be found in the last section of the paper.

Theorem 4.2. *For the SPA model, the logarithmic behaviour of the expected value of $e(\alpha)$ is as follows.*

For $1/2 < pA_1 < 1$,

$$\frac{\log \mathbb{E}(e(\alpha))}{\log n} = \begin{cases} 1 + o(1) & \text{if } \alpha \geq 1, \\ 2 - \frac{1}{pA_1} + \alpha \frac{1-pA_1}{pA_1} + o(1), & \text{if } 1 - pA_1 < \alpha < 1, \\ \frac{\alpha pA_1}{1-pA_1} + o(1), & \text{if } 0 \leq \alpha \leq 1 - pA_1. \end{cases}$$

For $pA_1 < 1/2$,

$$\frac{\log \mathbb{E}(e(\alpha))}{\log n} = \begin{cases} 1 + o(1) & \text{if } \alpha \geq 1, \\ \alpha + o(1), & \text{if } 0 \leq \alpha < 1. \end{cases}$$

Thus, for the case where $pA_1 > 1/2$, the middle range of $e(\alpha)$ extends beyond the lower bound on α for which precise results for $e(\alpha)$ can be obtained, and there is a third range for small α , namely $\alpha < 1 - pA_1$, for which the expected relationship between $\log e(\alpha)$ and α is given by

$$\log e(\alpha) = (1 + o(1)) \left(\frac{pA_1}{1 - pA_1} \right) \alpha \log n = -\frac{mpA_1}{1 - pA_1} \log r_\alpha. \quad (10)$$

Thus we have clear power law behaviour at the tail of the distribution, with coefficient $\frac{mpA_1}{1-pA_1} > 1$.

To verify our intuition that the real behaviour of the SPA model is similar to the asymptotic results given by the theorems, we ran simulations. We generated graphs of 100,000 nodes, in S of dimension $m = 2$, for various values of p . A_1 and A_2 were both set to 1. The results are seen in Figures 5 and 6, where the logarithm of the number of long edges has been plotted against a range of values for α . The straight lines in the figures represent the expected behaviour for the three ranges of α as given by (9) and (10) (and a horizontal line for the behaviour for large α). To show the fact that the number of long edges decreases as the threshold r_α increases, the x -axis gives the values of $-\alpha$.

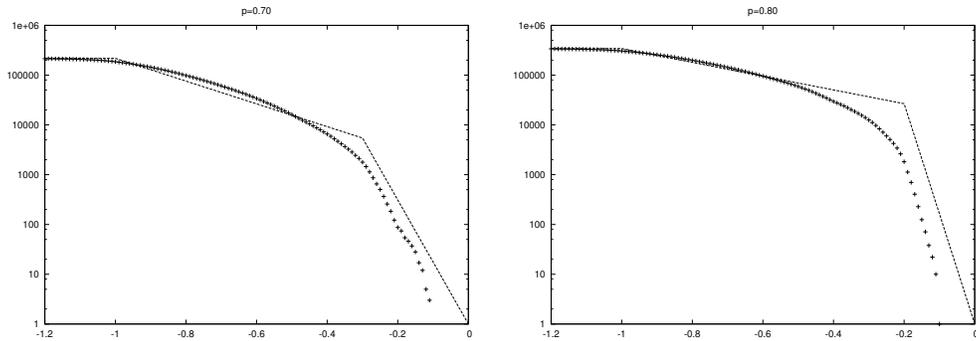


FIGURE 5. Long Edges Simulation vs. Theory, SPA Model. Parameters: $n = 100,000$, $A_1 = A_2 = 1$, $m = 2$, $p = 0.7$ (left) and $p = 0.8$ (right).

Figure 5 shows two values in the range $1/2 < pA_1 < 1$. For both cases, the theoretical results expressed in Equations 9 and 10 give a good approximation of the envelope of the curve represented by the simulated values. Not surprisingly, near the threshold $1 - pA_1 = \alpha$, the simulated version shows smooth behaviour that is a blend between the behaviour on both sides of the range. The angle of the tail of the distribution has good agreement with the value predicted from the modified model.

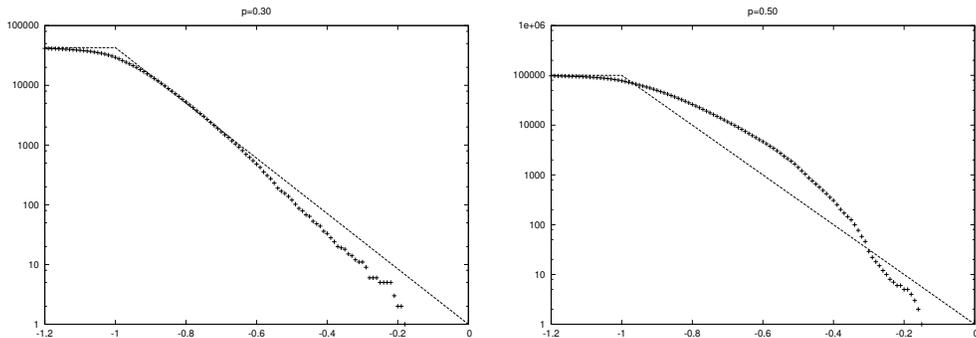


FIGURE 6. Long Edges Simulation vs. Theory, SPA Model. Parameters: $n = 100,000$, $A_1 = A_2 = 1$, $m = 2$, $p = 0.3$ (left), and $p = 0.5$ (right)

In Figure 6 we give a simulation result for the case where $pA_1 < 1/2$. Here the modified model predicts only two regimes, which is borne out by the simulation data. We also include a picture for the case $pA_1 = 1/2$. At this cross-over value, no linear relationship between $\log e(\alpha)$ and α can be observed from the picture. However, our theoretical results predict that for larger values of n , the curve should approach a straight line with slope $-\alpha$.

5. PROOFS OF THE MAIN THEOREMS

5.1. Degree of a vertex. The first part of this section is devoted the proof of Theorem 2.2. We will be using the following version of a well-known Bernstein inequalities many times so let us state it explicitly.

Lemma 5.1 ([13]). *Let X be a random variable that can be expressed as a sum $X = \sum_{i=1}^n X_i$ of independent random indicator variables where $X_i \in \text{Be}(p_i)$ with (possibly) different $p_i = \mathbb{P}(X_i = 1) = \mathbb{E}X_i$. Then the following holds for $t \geq 0$:*

$$\begin{aligned} \mathbb{P}(X \geq \mathbb{E}X + t) &\leq \exp\left(-\frac{t^2}{2(\mathbb{E}X + t/3)}\right), \\ \mathbb{P}(X \leq \mathbb{E}X - t) &\leq \exp\left(-\frac{t^2}{2\mathbb{E}X}\right). \end{aligned}$$

In particular, if $\varepsilon \leq 3/2$, then

$$\mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon \mathbb{E}X) \leq 2 \exp\left(-\frac{\varepsilon^2 \mathbb{E}X}{3}\right). \quad (11)$$

Now, we are ready to prove the following key observation.

Theorem 5.2. *Suppose that $\deg^-(v, T) = d \geq \omega \log n$, where $\omega = \omega(n)$ is any function tending to infinity together with n . Then, for every value of t , $T \leq t \leq 2T$, we get that*

$$\left| \deg^-(v, t) - d \cdot \left(\frac{t}{T}\right)^{pA_1} \right| \leq \frac{2}{pA_1} \cdot \frac{t}{T} \sqrt{d \log n}$$

with probability $1 - O(n^{-4/3})$.

Proof. Let $\omega = \omega(n)$ be any function tending to infinity together with n . Suppose that $\deg^-(v, T) = d \geq \omega \log n$. We will show that the upper bound holds; the lower bound can be obtained by using an analogous symmetric argument.

Let us introduce the following stopping time

$$T_0 = \min \left\{ t \geq T : \deg^-(v, t) > d \cdot \left(\frac{t}{T}\right)^{pA_1} + \frac{2}{pA_1} \cdot \frac{t}{T} \sqrt{d \log n} \vee t = 2T + 1 \right\}.$$

A *stopping time* is any random variable T_0 with values in $\{0, 1, \dots\} \cup \{\infty\}$ such that it can be determined whether $T_0 = t^*$ for any time t^* from knowledge of the process up to and including time t^* . The name can be misleading, since a process does not *stop* when it reaches a stopping time. Here, T_0 determines the first time the process does *not* exhibit the bounded behaviour we wish to establish. The

condition $t = 2T + 1$ has been added to assure that the set is never empty, and thus T_0 is well-defined. If $T_0 = 2T + 1$, then the in-degree of v remained bounded as given during the entire time interval $T \leq t \leq 2T$. In order to prove the bound, we need to show that with probability $1 - O(n^{-4/3})$ we have $T_0 = 2T + 1$.

Suppose that $T_0 \leq 2T$. Note that for $t \geq T$ up to and including time-step $T_0 - 1$, the random variable $\deg^-(v, t)$ is (deterministically) bounded from above, and so the number of new neighbours accumulated during this phase of the process, $\deg^-(v, T_0) - \deg^-(v, T)$, can be (stochastically) bounded from above by the sum $X = \sum_{t=T}^{T_0-1} X_t$ of independent indicator random variables X_t with

$$\mathbb{P}(X_t = 1) = p \frac{A_1 \left(d \left(\frac{t}{T} \right)^{pA_1} + \frac{2}{pA_1} \cdot \frac{t}{T} \sqrt{d \log n} \right) + A_2}{t}.$$

Hence,

$$\begin{aligned} \mathbb{E} \deg^-(v, T_0) &\leq d + \mathbb{E}X = d + \sum_{t=T}^{T_0-1} \mathbb{E}X_t \\ &= d + pA_1 d T^{-pA_1} \left(\sum_{t=T}^{T_0-1} t^{pA_1-1} \right) + \frac{T_0 - T}{T} 2\sqrt{d \log n} + O(1) \\ &= d \left(\frac{T_0}{T} \right)^{pA_1} + \frac{T_0 - T}{T} 2\sqrt{d \log n} + O(1). \end{aligned}$$

This implies that

$$\begin{aligned} \deg^-(v, T_0) - \mathbb{E} \deg^-(v, T_0) &\geq \frac{2}{pA_1} \cdot \frac{T_0}{T} \sqrt{d \log n} - \frac{T_0 - T}{T} 2\sqrt{d \log n} - O(1) \\ &\geq 2\sqrt{d \log n}, \end{aligned}$$

and it follows from the bound (11) that

$$\mathbb{P}(|X - \mathbb{E}X| \geq 2\sqrt{d \log n}) \leq 2 \exp \left(-\varepsilon \frac{2\sqrt{d \log n}}{3} \right),$$

where $\varepsilon = 2\sqrt{d \log n}/\mathbb{E}X$. Since the maximum value of $\mathbb{E}X$ corresponds to $T_0 = 2T$, it follows that $\mathbb{E}X \leq d(2^{pA_1} - 1)(1 + o(1)) \leq d$, and so $\varepsilon \geq 2\sqrt{d^{-1} \log n}$. Therefore, the probability that $T_0 \leq 2T$ is at most $2 \exp(-\frac{4}{3} \log n)$ and the theorem is finished. \square

Now, with Theorem 5.2 in hand we can easily get Theorem 2.2. For a given vertex v of degree $\omega \log n$ at time T we obtain from Theorem 5.2 that, with probability $1 - O(n^{-4/3})$,

$$d \left(\frac{t}{T} \right)^{pA_1} \left(1 - \frac{4}{pA_1} \sqrt{d^{-1} \log n} \right) \leq \deg^-(v, t) \leq d \left(\frac{t}{T} \right)^{pA_1} \left(1 + \frac{4}{pA_1} \sqrt{d^{-1} \log n} \right)$$

for $T \leq t \leq 2T$. We can now keep applying the same theorem for times $2T, 4T, 8T, 16T, \dots$, using the final value as the initial one for the next period, to get the statement for all values of t from T up to and including time n . Since we apply the theorem $O(\log n)$ times (for a given vertex v), the statement holds with

probability $1 - o(n^{-1})$ and so *a.a.s.* the statement we are about to prove will hold for all vertices.

It remains to make sure that the accumulated multiplicative error is still only $(1 + o(1))$. After applying the theorem recursively i times the degree is shown to be $d2^{pA_1 i}(1 + o(1))$. Using this rough estimate, and assuming the theorem is applied for a total of $k = O(\log n)$ times, we get that the error term is, in fact, bounded from above by

$$\begin{aligned} \prod_{i=1}^k \left(1 + \frac{5}{pA_1} \sqrt{d^{-1}2^{-pA_1 i} \log n} \right) &= (1 + o(1)) \exp \left(\frac{5}{pA_1} \sqrt{d^{-1} \log n} \sum_{i=1}^k 2^{-pA_1 i/2} \right) \\ &= (1 + o(1)) \exp \left(O(\sqrt{d^{-1} \log n}) \right) \\ &= 1 + o(1), \end{aligned}$$

since d grows faster than $\log n$. A symmetric argument can be used to show a lower bound for the error term and so Theorem 2.2 holds.

5.2. Number of common neighbours. The proof of Theorem 3.1, which gives a formula for the number of common neighbours of two given vertices v and w , is based on three cases, as explained in Section 3 and Figure 1. The division into three cases is based on the trend, as shown in Theorem 2.2, that spheres of influence tend to shrink over time. It can happen that spheres of influence that are disjoint become overlapping at a later time instance, and thus do not fit any of the three cases. However, this behaviour happens with low enough probability that it does not affect our result.

Proof of Theorem 3.1. The proof depends heavily on Theorem 2.2. Any precise reference to the theorem will therefore be omitted. We can assume that at time T ,

$$\deg(v_\ell, T) = (1 + o(1)) \frac{A_2}{A_1} \omega \log n$$

and the degree of this vertex is as predicted by (4) until the end of the process (that is, the ratio between the upper and lower bound on the degree is deterministically equal to $(1 + o(1))$). Since $k \geq l$, the degree of v_k for the time interval after T is given by (4) as well. Let $r(v, t)$ denote the radius of the sphere of influence around v at time t ; that is, $r(v, t) = (|S(v, t)|/c_m)^{1/m}$.

Case 1: Suppose that $d \geq \varepsilon(\omega \log n/T)^{1/m}$ for some $\varepsilon > 0$. For $T \leq t \leq n$, we can deduce from the expression for the degree of v_ℓ over time and the expression for the volume $|S(v_\ell, t)|$ of the sphere of influence of v_ℓ that

$$r(v_\ell, t) = (1 + o(1)) \left(\frac{A_2 \omega \log n (t/T)^{pA_1}}{c_m t} \right)^{1/m}.$$

In particular, let us note that d is of greater or equal order as $r(v_\ell, T)$, and hence of greater or equal order as $r(v_k, T)$ as well. Moreover, both radii tend to be decreasing from time T on. (Formally what we mean is that $r(v_\ell, t) > r(v_\ell, t(1+\varepsilon))$ for any $\varepsilon > 0$ and $t \geq T$. When a vertex receives a new neighbour, its radius slightly increases.) Therefore, there exists a constant $c = c(\varepsilon) > 0$ such that

$S(v_\ell, t)$ and $S(v_k, t)$ are disconnected for every $t > cT$ and so there is no chance to create more common neighbours. Since at time cT the degree of vertex v_ℓ is $(1 + o(1))(A_2/A_1)c^{pA_1}\omega \log n = O(\omega \log n)$, we can apply an obvious upper bound to get

$$cn(v_\ell, v_k, n) \leq \deg(v_\ell, n) = O(\omega \log n).$$

Finally, note that it can happen that $cT > n$, which means that the process stops before the spheres of influence become disjoint. This causes no problem since the upper bound for the number of common neighbours at time cT will then trivially hold at time n .

Case 2: Suppose $k \geq (1 + \varepsilon)\ell$ for some $\varepsilon > 0$ and d satisfies inequality (5). Note that the condition for d implies that at time n the sphere of influence of v_ℓ is contained in that of v_k . Moreover, the radii of influence are proportionally decreasing during the process from the time we start having concentrated behaviour of degrees onwards (that is, from time T on, in the sense explained earlier). So the sphere of influence of v_ℓ is contained in the sphere of influence of v_k from time T to time $(1 + o(1))n$. Any vertex u that links to v_ℓ lies inside the sphere of influence of v_ℓ and thus of v_k as well, and has a probability p of also linking to v_k .

At the end of the process (for $t = (1 + o(1))n$) it can happen that the sphere of influence $S(v_\ell, t)$ is not completely contained in $S(v_k, t)$, but it is the case that they overlap to a large extent, namely

$$\frac{|S(v_\ell, t) \cap S(v_k, t)|}{|S(v_\ell, t)|} = 1 + o(1). \quad (12)$$

Thus, the probability that a neighbour of v_ℓ , added during this phase of the process, is also a neighbour of v_k is $(1 + o(1))p$.

Therefore, $\mathbb{E}cn(v_\ell, v_k, n) = (1 + o(1))p\ell$, since the number of common neighbours accumulated until time T is $O(\omega \log n)$ and so is negligible.

Suppose now that $k = (1 + o(1))\ell$ and $d \ll (k/n)^{1/m}$. In this case, the radii of v_ℓ and v_k are approximately equal from time T to the end of the process (that is, they differ by a multiplicative factor of $(1 + o(1))$). Since d is of order smaller than the radii at the end of the process, property (12) holds for $T \leq t \leq n$ and the results holds by the same argument as before.

Case 3: Suppose $k \geq (1 + \varepsilon)\ell$ for some $\varepsilon > 0$ and d satisfies inequality (6). Note that the condition for d implies that at time T the sphere of influence of v_ℓ is contained in that of v_k , but this is not the case at time n .

Let t^- be the first moment when $S(v_\ell, t)$ is not completely contained in $S(v_k, t)$ ($T < t^- \leq n$). Let t^+ be the last time when the spheres overlap ($t^- \leq t^+$). (Note that it is possible that $t^+ > n$ but, as before, this causes no problem.) Up to time t^- , each neighbour of v_ℓ will be a common neighbour of v_ℓ and v_k with probability p . From time t^+ to n , no common neighbours can be created. From time t^- until time t^+ , the probability that a neighbour of v_ℓ becomes a neighbour of v_k is *at most* p . Thus, $p \deg^-(v_\ell, t^-)$ and $p \deg^-(v_\ell, t^+)$ form a lower and an upper bound, respectively, on the expected number of common neighbours of v and w .

Note that at time t^- , $S(v_\ell, t^-)$ is contained in $S(v_k, t^-)$ and “touches” the boundary from the inside (the distance between the boundaries at time t^- may not be exactly zero but certainly is $o(d)$). At time t^+ , $S(v_\ell, t^+)$ is outside $S(v_k, t^-)$ but

“touches” the boundary from the outside. Since the centers of $S(v_\ell, t)$ and $S(v_k, t)$ are at distance d from each other, this translates into the following expressions involving t^- and t^+ :

$$\begin{aligned} r(v_k, t^-) - r(v_\ell, t^-) &= (1 + o(1))d, \\ r(v_k, t^+) + r(v_\ell, t^+) &= (1 + o(1))d. \end{aligned}$$

Using the concentration result about the in-degree, this translates into the following conditions on t^- and t^+

$$\begin{aligned} \left(\frac{A_2}{c_m}(t^-)^{pA_1-1}\right)^{1/m} i_k^{-pA_1/m} \left(1 - \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right) &= (1 + o(1))d, \\ \left(\frac{A_2}{c_m}(t^+)^{pA_1-1}\right)^{1/m} i_k^{-pA_1/m} \left(1 + \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right) &= (1 + o(1))d, \end{aligned}$$

and so

$$\begin{aligned} t^- &= (1 + o(1)) \left(\frac{A_2}{c_m}\right)^{\frac{1}{1-pA_1}} i_k^{-\frac{pA_1}{1-pA_1}} d^{-\frac{m}{1-pA_1}} \left(1 - \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right)^{\frac{m}{1-pA_1}}, \\ t^+ &= (1 + o(1)) \left(\frac{A_2}{c_m}\right)^{\frac{1}{1-pA_1}} i_k^{-\frac{pA_1}{1-pA_1}} d^{-\frac{m}{1-pA_1}} \left(1 + \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right)^{\frac{m}{1-pA_1}}. \end{aligned}$$

The number of common neighbours of v_k and v_ℓ is bounded from below by $(1 + o(1))p \deg^-(v_\ell, t^-)$, and from above by $(1 + o(1))p \deg(v_\ell, t^+)$. Using our knowledge about the behaviour of the in-degree of v_ℓ , this leads to the following bounds, which hold within a $(1 + o(1))$ term:

$$\begin{aligned} pA_1^{-1} A_2^{\frac{1}{1-pA_1}} c_m^{-\frac{pA_1}{1-pA_1}} i_k^{-\frac{(pA_1)^2}{1-pA_1}} i_\ell^{-pA_1} d^{-\frac{mpA_1}{1-pA_1}} \left(1 - \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right)^{\frac{mpA_1}{1-pA_1}} \\ \leq \mathbb{E} cn(v_\ell, v_k, n) \leq \\ pA_1^{-1} A_2^{\frac{1}{1-pA_1}} c_m^{-\frac{pA_1}{1-pA_1}} i_k^{-\frac{(pA_1)^2}{1-pA_1}} i_\ell^{-pA_1} d^{-\frac{mpA_1}{1-pA_1}} \left(1 + \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right)^{\frac{mpA_1}{1-pA_1}}. \end{aligned}$$

The result follows from the fact that

$$\left(1 \pm \left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right)^{\frac{mpA_1}{1-pA_1}} = 1 + O\left(\left(\frac{i_k}{i_\ell}\right)^{pA_1/m}\right).$$

Finally, consider the case where $k = (1 + o(1))\ell$, and thus $i_k/i_\ell = 1 + o(1)$. As before, from time

$$t^+ = (1 + o(1)) \left(\frac{A_2 2^m}{c_m}\right)^{\frac{1}{1-pA_1}} i_k^{-\frac{pA_1}{1-pA_1}} d^{-\frac{m}{1-pA_1}}$$

until time n , the spheres are disjoint and there is no chance for a common neighbour. At time t such that $T \leq t = o(t^+)$, the spheres overlap to a large extent and (12) holds. However, for $\varepsilon > 0$ and t such that $\varepsilon t^+ \leq t \leq t^+$ only a nontrivial

fraction of $S(v_\ell, t)$ is contained in $S(v_k, t)$. The above analysis still applies, but in this case instead of an asymptotic result, we obtain the order result stated in the theorem.

Finally, let us note that the number of common neighbours is a sum of independent random indicator variables with Bernoulli distribution. The concentration follows from the bound (5.1). \square

5.3. Edge length distribution. Finally, we give the proof of the theorem about the edge length distribution. Remember that a long edge is an edge such that its endpoints are at distance at least r_α , where r_α is chosen so that a ball of radius r_α has volume $n^{-\alpha}$. As in the previous subsection, the proof distinguishes three cases, but now the three cases depend on whether the sphere of influence of a vertex has radius greater than r_α (allowing the vertex to receive long edges) at the beginning and the end of its life.

First, we need to recall a few known results: the behaviour of $N_k = N_k(n)$, the number of vertices of in-degree $k = k(n)$ at time n , the number of edges $M = M(n)$ at time n , and the upper bound for the size of the influence regions. The following result was proven in [1].

Theorem 5.3 ([1]). *Suppose that $pA_1 < 1$. The following holds a.a.s. for every $0 \leq k \leq (n/\log^8 n)^{(pA_1)/(4pA_1+2)}$.*

$$N_k = (1 + o(1))c_k n,$$

where $c_0 = 1/(1 + pA_2)$ and for $k \geq 1$,

$$c_k = \frac{p^k}{1 + kpA_1 + pA_2} \prod_{j=0}^{k-1} \frac{jA_1 + A_2}{1 + jpA_1 + pA_2}.$$

Moreover, a.a.s.

$$M = (1 + o(1)) \frac{pA_2}{1 - pA_1} n.$$

Note that

$$\begin{aligned} c_k &= \frac{1}{pA_1} \cdot \frac{\prod_{j=0}^{k-1} \left(j + \frac{A_2}{A_1}\right)}{\prod_{j=0}^k \left(j + \frac{A_2}{A_1} + \frac{1}{pA_1}\right)} \\ &= \frac{1}{pA_1} \cdot \frac{\Gamma\left(k + \frac{A_2}{A_1}\right) / \Gamma\left(\frac{A_2}{A_1}\right)}{\Gamma\left(k + 1 + \frac{A_2}{A_1} + \frac{1}{pA_1}\right) / \Gamma\left(\frac{A_2}{A_1} + \frac{1}{pA_1}\right)}. \end{aligned}$$

Suppose now that $k = k(n)$ tends to infinity together with n . Using Stirling's asymptotic approximation of the Gamma function ($\Gamma(z) = (1 + o(1))\sqrt{2\pi}z^{z-1/2}e^{-z}$) we can take c_k to be:

$$c_k = \frac{1}{pA_1} \cdot \frac{\Gamma\left(\frac{A_2}{A_1} + \frac{1}{pA_1}\right)}{\Gamma\left(\frac{A_2}{A_1}\right)} k^{-1 - \frac{1}{pA_1}},$$

and the following useful corollary is proved.

Corollary 5.4. *Suppose that $pA_1 < 1$. Let $\omega = \omega(n)$ be any function tending to infinity with n . The following holds a.a.s. for every $\omega \leq k \leq (n/\log^8 n)^{(pA_1)/(4pA_1+2)}$.*

$$N_k = (1 + o(1))ck^{-1-\frac{1}{pA_1}}n,$$

where

$$c = \frac{1}{pA_1} \cdot \frac{\Gamma\left(\frac{A_2}{A_1} + \frac{1}{pA_1}\right)}{\Gamma\left(\frac{A_2}{A_1}\right)}. \quad (13)$$

In [1], it was proved that a.a.s. for all vertices we have that $\deg^-(v_i, n) = O((\log^2 n)(n/i)^{pA_1})$, provided that v_i was born at time i . Now, with Theorem 2.2 in hand, we get a stronger result, namely that a.a.s. for all $i \leq t \leq n$

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

where $\omega = \omega(n)$ is any function tending to infinity together with n . (Indeed, for a contradiction suppose that $\deg^-(v_i, t) \geq (2\omega \log n)(t/i)^{pA_1}$ for some value of t . Theorem 2.2 implies that $\deg^-(v_i, i) = (2 + o(1))\omega \log n$ which is clearly a contradiction.) This implies the following result.

Theorem 5.5. *Suppose that $pA_1 < 1$, and ω is a function that goes to infinity together with n . The following holds a.a.s. for every vertex born at time i .*

$$|S(v_i, t)| = O\left(\frac{\omega \log n}{i}\right),$$

The results given above are used in the proof of Theorem 4.1, which we are now ready to give.

Proof of Theorem 4.1. Suppose first that $\alpha > 1$. Since the sphere of influence of every vertex at every time of the process is (deterministically) at least $A_2/n \gg n^{-\alpha}$, “long” edges can occur at every step of the process. A vertex v will receive a short edge precisely when the new vertex falls within a ball of radius r_α around v , and thus automatically falls within the sphere of influence of v , and then links to v . The probability that this happens equals $pn^{-\alpha}$. Thus, the expected number of short edges pointing to a vertex born at time i is $pn^{-\alpha}(n-i)$, and the total number of short edges is $(1 + o(1))pn^{2-\alpha}/2 = o(n)$ and so is negligible compared to the total number of edges. We conclude that a.a.s. almost all edges are long, and the result holds by Theorem 5.3.

Suppose now that $1 - \frac{pA_1}{4pA_1+2} < \alpha < 1$. Let $e_v(\alpha)$ be the number of long edges pointing to v , that is:

$$e_v(\alpha) = \left| \{w \in N^-(v) : d(v, w) \geq r_\alpha\} \right|,$$

where $N^-(v)$ is the in-neighbourhood of vertex v .

For a vertex v to receive an edge of length greater than r_α at time t , its region of influence must have radius at least r_α , and thus have volume $|S(v, t)| \geq n^{-\alpha}$.

Key to the proof is Theorem 2.2 and its conclusion that the regions of influence tend to be shrinking.

Let $\omega = \omega(n)$ be any function increasing with n . First, we only consider vertices whose final degree is at least $\omega \log n$. This is enough to get a lower bound for the number of long edges. Later we will show that the contribution of the remaining edges is negligible. Consider a vertex v with final degree $k = \deg^-(v, n) \geq \omega \log n$. It follows from Theorem 2.2 that *a.a.s.* for every vertex v of degree $k \geq \omega \log n$ at time n ,

$$\deg^-(v, t) = (1 + o(1))k \left(\frac{t}{n}\right)^{pA_1}$$

for all $t_k \leq t \leq n$, where

$$t_k = n \left(\frac{\omega \log n}{k}\right)^{\frac{1}{pA_1}}.$$

(Note that $\deg(v, t_k) = (1 + o(1))\omega \log n$.) Therefore, we may assume, without loss of generality, that for all $t_k \leq t \leq n$

$$|S(v, t)| = (1 + o(1))A_1 k n^{-pA_1} t^{pA_1 - 1}.$$

We distinguish three possible classes of vertices, based on their final degree: vertices of high final degree can receive long edges from time t_k until the end of the process, $t = n$ (Case 1); vertices with final degree in a mid-range can receive long edges from time t_k until some time t_k^* , $t_k < t_k^* < n$ (Case 2); and vertices with small final degree can never receive long edges after time t_k (Case 3).

The cut-off values of the three cases are

$$k_{\min} = \left(\frac{n^{1-\alpha}}{A_1}\right)^{pA_1} (\omega \log n)^{1-pA_1}$$

and $k_{\max} = \frac{n^{1-\alpha}}{A_1}$. Consider a vertex v of degree k .

Case 1. Suppose that $k \geq k_{\max}$. Note that this implies that

$$|S(v, n)| = (1 + o(1))A_1 k/n \geq (1 + o(1))n^{-\alpha},$$

so for any constant $\varepsilon > 0$, and for any time t in the range $t_k \leq t \leq (1 - \varepsilon)n$, the sphere of influence of v has radius greater than r_α . This implies that v has an opportunity to receive long edges from time t_k until the end of the process, or very close to it.

For $t_k \leq t \leq n$, the probability that v receives a short edge (edge from a vertex within distance r_α) equals $p \min\{n^{-\alpha}, |S(v, t)|\} = (1 + o(1))pn^{-\alpha}$. Moreover, these events are independent. Thus, *w.e.p.* the number of short edges is

$$(1 + o(1))pn^{-\alpha}(n - t_k) = (1 + o(1))pn^{1-\alpha},$$

where the last step uses the fact that $t_k = o(n)$ in this case.

The degree of v at time t_k is $O(\omega \log n)$, so we have that *w.e.p.*

$$\begin{aligned} e_v(\alpha) &= \deg^-(v, n) - (1 + o(1))pn^{1-\alpha} + O(\omega \log n) = (1 + o(1))(k - pn^{1-\alpha}) \\ &\geq (1 - pA_1 + o(1))\frac{n^{1-\alpha}}{A_1}. \end{aligned}$$

Note that if $k \geq \omega n^{1-\alpha}$, then *w.e.p.* almost all edges pointing to v are long.

Case 2. Let $\varepsilon > 0$ be some (arbitrarily small) constant. Suppose that $(1 + \varepsilon)k_{\min} \leq k \leq (1 - \varepsilon)k_{\max}$. The upper bound on k implies that $|S(v, n)| \leq (1 - \varepsilon + o(1))n^{-\alpha}$ so there is no chance for v to receive long edges near the end of the process. On the other hand, it follows from the lower bound on k that $|S(v, t_k)| \geq (1 + \varepsilon - o(1))n^{-\alpha}$ so if the new vertex at time t_k falls within $S(v, t_k)$, there is a positive probability that a long edge to v is created.

Let

$$t_k^* = (A_1 k n^{\alpha - pA_1})^{\frac{1}{1 - pA_1}}.$$

Note that $|S(v, t_k^*)| = (1 + o(1))n^{-\alpha}$. Thus, the influence region of v has radius greater than r_α from time t_k to $(1 - \delta)t_k^*$, and radius less than r_α from time $(1 + \delta)t_k^*$ to n , for some small $\delta > 0$.

Thus, by a similar argument to the previous case, we obtain:

$$\begin{aligned} e_v(\alpha) &\geq (1 + o(1)) \sum_{t=t_k}^{(1-\delta)t_k^*} p (A_1 k n^{-pA_1} t^{pA_1-1} - n^{-\alpha}) \\ &= (1 - O(\delta)) (k n^{-pA_1} (t_k^*)^{pA_1} - p(t_k^*)n^{-\alpha}) \\ &= (1 - O(\delta)) A_1^{\frac{pA_1}{1-pA_1}} k^{\frac{1}{1-pA_1}} n^{\frac{-pA_1(1-\alpha)}{1-pA_1}} (1 - pA_1). \end{aligned}$$

Similarly, we get that $e_v \leq (1 + O(\delta)) A_1^{\frac{pA_1}{1-pA_1}} k^{\frac{1}{1-pA_1}} n^{\frac{-pA_1(1-\alpha)}{1-pA_1}} (1 - pA_1)$ and so

$$e_v(\alpha) = (1 + o(1)) A_1^{\frac{pA_1}{1-pA_1}} k^{\frac{1}{1-pA_1}} n^{\frac{-pA_1(1-\alpha)}{1-pA_1}} (1 - pA_1),$$

by taking $\delta \rightarrow 0$.

Case 3. Finally, suppose that $\omega \log n \leq k \leq (1 - \varepsilon)k_{\min}$ for some $\varepsilon > 0$. Since $|S(v, t_k)| \leq (1 - \varepsilon + o(1))n^{-\alpha}$, the influence region has radius smaller than r_α from time t_k until the end of the process. Thus, for such vertices, all edges they receive in this time slot are short. Thus the only long edges v can receive are those received before t_k^* , so $e_v(\alpha) = O(\omega \log n)$. Trivially, the same property holds for any vertex of degree smaller than $\omega \log n$.

In order to obtain upper and lower bounds on the total number of long edges, we can use Theorem 5.3 and its corollary (Corollary 5.4) to calculate the number of long edges pointing to vertices of final degree larger than k_{\min} (Cases 1 and 2). Let c be as defined in Equation (13), and let $K = (n / \log^8 n)^{(pA_1)/(4pA_1+2)}$. By Corollary 5.4, K is the upper bound on the values of k for which we have concentration for N_k . Note that the bounds on α imply that $k_{\max} \ll K$, and thus $\sum_{k \geq k_{\max}} k^{-\gamma} = (1 + o(1)) \sum_{k_{\max} \leq k \leq K} k^{-\gamma}$ for all $\gamma > 1$.

The number of long edges to vertices of the first type (Case 1) is *a.a.s.* equal to

$$\begin{aligned}
E_1 &= (1 + o(1)) \sum_{k \geq k_{\max}} N_k (k - pn^{1-\alpha}) \\
&= (1 + o(1)) \sum_{k_{\max} \leq k \leq K} \left(ck^{-1-\frac{1}{pA_1}} n \right) (k - pn^{1-\alpha}) \\
&= (1 + o(1)) \left(cn \sum_{k \geq k_{\max}} k^{-\frac{1}{pA_1}} - cpn^{2-\alpha} \sum_{k \geq k_{\max}} k^{-1-\frac{1}{pA_1}} \right) \\
&= (1 + o(1)) \left(cn \frac{(k_{\max})^{\frac{pA_1-1}{pA_1}}}{\frac{1-pA_1}{pA_1}} - cpn^{2-\alpha} \frac{(k_{\max})^{-\frac{1}{pA_1}}}{\frac{1}{pA_1}} \right) \\
&= c \frac{pA_1^{\frac{1}{pA_1}}}{1-pA_1} n^{2-\frac{1}{pA_1}+\alpha\frac{1-pA_1}{pA_1}} (1 - (pA_1)(1-pA_1)).
\end{aligned}$$

The number of long edges to vertices of the second type (Case 2) is *a.a.s.* equal to

$$\begin{aligned}
E_2 &= (1 + o(1)) \sum_{k=k_{\min}}^{k_{\max}} \left(ck^{-1-\frac{1}{pA_1}} n \right) \left(A_1^{\frac{pA_1}{1-pA_1}} k^{\frac{1}{1-pA_1}} n^{-\frac{pA_1(1-\alpha)}{1-pA_1}} (1-pA_1) \right) \\
&= (1 + o(1)) c A_1^{\frac{pA_1}{1-pA_1}} (1-pA_1) n^{1-\frac{pA_1(1-\alpha)}{1-pA_1}} \sum_{k=k_{\min}}^{k_{\max}} k^{-1+\frac{2pA_1-1}{(1-pA_1)pA_1}}.
\end{aligned}$$

(Technically, to get a lower bound of E_2 we should sum over $k_{\min}(1+\varepsilon) \leq k \leq k_{\max}(1-\varepsilon)$ and sum over $k_{\min}(1+\varepsilon) \leq k \leq k_{\max}$ to get an upper bound. Since the error in this summation is $(1+O(\varepsilon))$, the result holds by taking $\varepsilon \rightarrow 0$.)

Since $1/2 < pA_1 < 1$, the exponent of k in the summation is in the interval $(-1, 0)$, and thus the behaviour of the summation is determined by its upper bound k_{\max} . This leads to

$$\begin{aligned}
E_2 &= (1 + o(1)) c A_1^{\frac{pA_1}{1-pA_1}} (1-pA_1) n^{1-\frac{pA_1(1-\alpha)}{1-pA_1}} \frac{(k_{\max})^{\frac{2pA_1-1}{(1-pA_1)pA_1}}}{\frac{2pA_1-1}{(1-pA_1)pA_1}} \\
&= (1 + o(1)) c \frac{pA_1^{\frac{1}{pA_1}}}{1-pA_1} n^{2-\frac{1}{pA_1}+\alpha\frac{1-pA_1}{pA_1}} \cdot \frac{(1-pA_1)^3}{2pA_1-1} A_1^{\frac{1-2pA_1}{(1-pA_1)pA_1}}.
\end{aligned}$$

Since E_1 and E_2 are of the same order, we can take $E_1 + E_2$ as a lower bound for $e(\alpha)$.

In order to obtain an upper bound, we consider edges to vertices that are in Case 3, that is, those that have final degree at most k_{\min} . It follows from Theorem 5.5 that any vertex that is able to receive long edges directed to vertices with small final degree has to have a time of birth $i \leq i_{\max} = \omega n^\alpha \log n$. There are obviously at most i_{\max} of such vertices, and each of them has $O(\omega \log n)$ long edges. So the number of long edges that we did not count yet is at most:

$$E_3 = O((\omega n^\alpha \log n)(\omega \log n)) = n^{\alpha+o(1)}.$$

Since E_3 is of smaller order than $E_1 + E_2$, the result follows. \square

For the proof of Theorem 4.2, we use a large part of the previous proof.

Proof of Theorem 4.2. For this theorem, we consider the expected value of $e(\alpha)$. Thus, we can use the expected values of N_k , and do not need to consider the cut-off on the values of k for which the values of N_k are concentrated. In [1], it was shown that

$$\mathbb{E}(N_k) = (1 + o(1))ck^{-1-\frac{1}{pA_1}}n, \text{ for all } k \geq k_{\max}.$$

Suppose first that $1/2 < pA_1 < 1$ and $1 - pA_1 < \alpha < 1$. Consider the proof of Theorem 4.1. The three cases of this proof still hold as before; let k_{\min} and k_{\max} be as defined in this proof. As explained in this proof, concentration for the values of N_k hold only up to degree $K = n^{pA_1/(4pA_1+2)+o(1)}$. This affects the computation of E_1 . However, in this proof we only consider the expected value of $e(\alpha)$, so by linearity of expectation, in the computation of E_1 we can use the expected values of the N_k . This leads to the following expression for the expected number of long edges to vertices of the first type (Case 1) :

$$\begin{aligned} \mathbb{E}(E_1) &= (1 + o(1)) \sum_{k \geq k_{\max}} \mathbb{E}(N_k) (k - pn^{1-\alpha}) \\ &= (1 + o(1)) \sum_{k \geq k_{\max}} \left(ck^{-1-\frac{1}{pA_1}}n \right) (k - pn^{1-\alpha}) \\ &= c \frac{pA_1^{\frac{1}{pA_1}}}{1 - pA_1} n^{2-\frac{1}{pA_1}+\alpha\frac{1-pA_1}{pA_1}} (1 - (pA_1)(1 - pA_1)), \end{aligned} \quad (14)$$

where c is as defined in Equation (13).

For the computation of E_2 , we should note that we may not have concentration of N_k for the values of k close to k_{\max} . However, we can a similar calculation to that used in the proof of Theorem 4.1, using the expected values of the N_k , to obtain that $\mathbb{E}(E_2) = \Theta(n^{2-\frac{1}{pA_1}+\alpha\frac{1-pA_1}{pA_1}})$.

The argument that E_3 is negligible compared to $\mathbb{E}(E_1)$ and $\mathbb{E}(E_2)$, as laid out in the proof of Theorem 4.1, still holds here. Thus, we have that

$$\mathbb{E}(e(\alpha)) = \Theta(n^{2-\frac{1}{pA_1}+\alpha\frac{1-pA_1}{pA_1}})$$

The result follows by taking the logarithm.

Next, consider the case where $1/2 < pA_1 < 1$ and $\alpha < 1 - pA_1$. It follows from Theorem 2.2, and it was also shown in [1], that w.e.p. the maximum in-degree in a graph produced by the SPA model is at most $K_M = O(n^{pA_1} \log^4 n)$. Since $k_{\max} = n^{1-\alpha}/A_1 \gg n^{pA_1}$, w.e.p. no vertices are in Case 1, so no vertices can receive long edges until the edge of the process.

For the vertices that are in Case 2, we can apply the same calculation as in the proof of Theorem 4.1, while taking the expected values of the N_k as in the previous case. Since w.e.p. K_M is an upper bound on the maximum degree, the expected number of vertices of degree greater than K_M is at most $n \exp(-\Theta(\log^2 n))$. Hence, the expected number of long edges to such vertices,

is at most $n^2 \exp(-\Theta(\log^2 n)) = o(1)$. The expected number of long edges to vertices of the second type (Case 2) therefore is equal to

$$\begin{aligned} \mathbb{E}(E_2) &\leq (1 + o(1)) \sum_{k=k_{\min}}^{K_M} \left(ck^{-1-\frac{1}{pA_1}} n \right) \left(A_1^{\frac{pA_1}{1-pA_1}} k^{\frac{1}{1-pA_1}} n^{\frac{-pA_1(1-\alpha)}{1-pA_1}} (1 - pA_1) \right) \\ &= (1 + o(1)) c A_1^{\frac{pA_1}{1-pA_1}} (1 - pA_1) n^{1-\frac{pA_1(1-\alpha)}{1-pA_1}} \sum_{k=k_{\min}}^{K_M} k^{-1+\frac{2pA_1-1}{(1-pA_1)pA_1}}. \end{aligned}$$

For a lower bound on $\mathbb{E}(E_2)$, we should sum over $k_{\min}(1 + \varepsilon) \leq k \leq n^{pA_1}$.

Since $pA_1 > 1/2$, as before the exponent of k in the summation is determined by its upper bound $K_M = O(n^{pA_1} \log^4 n)$. This leads to

$$\begin{aligned} \mathbb{E}(E_2) &\leq (1 + o(1)) c A_1^{\frac{pA_1}{1-pA_1}} (1 - pA_1) n^{1-\frac{pA_1(1-\alpha)}{1-pA_1}} \frac{(K_M)^{\frac{2pA_1-1}{(1-pA_1)pA_1}}}{\frac{2pA_1-1}{(1-pA_1)pA_1}} \\ &= g(n) n^{\frac{pA_1\alpha}{1-pA_1}}, \end{aligned}$$

for some function g of order $g(n) = \Theta(K_M/n^{pA_1}) = \Theta(\log^4 n)$. For the lower bound on $\mathbb{E}(E_2)$, we have the same summation, but with upper bound n^{pA_1} instead of K_M , and thus $\mathbb{E}(E_2) = \Omega(n^{\frac{pA_1\alpha}{1-pA_1}})$.

Since $\log(g(n)) = o(\log n)$, we can combine lower and upper bound to see that

$$\frac{\log \mathbb{E}(E_2)}{\log n} = \frac{pA_1\alpha}{1-pA_1} + o(1).$$

Finally, consider the vertices in Case 3. Here, the exact same argument as given in the proof of Theorem 4.1 can be used to show that

$$\mathbb{E}(E_3) = O((\omega n^\alpha \log n)(\omega \log n)) = n^{\alpha+o(1)}.$$

Since this is of smaller order than $\mathbb{E}(E_2)$, the result follows.

Finally, consider the case where $pA_1 \leq 1/2$. For $\alpha \in (1 - pA_1, 1)$ we have the exact same expression for $\mathbb{E}(E_1)$ as for the case where $pA_1 > 1/2$, as given in Equation (14). Thus

$$\mathbb{E}(E_1) = \Theta(n^{2-\frac{1}{pA_1} + \alpha \frac{1-pA_1}{pA_1}}) = o(n^\alpha),$$

where the last step follows since

$$2 - \frac{1}{pA_1} + \alpha \frac{1-pA_1}{pA_1} = 1 - (1 - \alpha) \left(\frac{1-pA_1}{pA_1} \right) < \alpha.$$

For $\alpha \in (0, 1 - pA_1)$ we have that w.e.p. $E_1 = 0$, so $\mathbb{E}(E_1) = \exp(-\Theta(\log^2 n))$.

For E_2 , we have the same sum as before: let $K^* = k_{\max}$ if $\alpha \in (1 - pA_1, 1)$, and K^* , an (almost sure) upper bound on the maximum degree, otherwise. Then

$$\begin{aligned} \mathbb{E}(E_2) &= (1 + o(1)) \sum_{k=k_{\min}}^{K^*} \left(ck^{-1-\frac{1}{pA_1}} n \right) \left(A_1^{\frac{pA_1}{1-pA_1}} k^{\frac{1}{1-pA_1}} n^{-\frac{pA_1(1-\alpha)}{1-pA_1}} (1 - pA_1) \right) \\ &= (1 + o(1)) c A_1^{\frac{pA_1}{1-pA_1}} (1 - pA_1) n^{1-\frac{pA_1(1-\alpha)}{1-pA_1}} \sum_{k=k_{\min}}^{K^*} k^{-1+\frac{2pA_1-1}{(1-pA_1)pA_1}}. \end{aligned}$$

Since $pA_1 < 1/2$, the exponent of k in the summation in this case is determined by its lower bound

$$k_{\min} = \left(\frac{n^{1-\alpha}}{A_1} \right)^{pA_1} (\omega \log n)^{1-pA_1}.$$

This leads to

$$\mathbb{E}(E_2) \leq (1 + o(1)) c A_1^{\frac{pA_1}{1-pA_1}} (1 - pA_1) n^{1-\frac{pA_1(1-\alpha)}{1-pA_1}} \frac{(k_{\min})^{\frac{2pA_1-1}{(1-pA_1)pA_1}}}{\frac{2pA_1-1}{(1-pA_1)pA_1}} = o(n^\alpha),$$

where the last step follows since the exponent of $(\omega \log n)$ in $(k_{\min})^{\frac{2pA_1-1}{(1-pA_1)pA_1}}$ equals

$$(1 - pA_1) \frac{(2pA_1 - 1)}{(1 - pA_1)pA_1} < 0.$$

Finally, the same estimate as before can be used to show that $E_3 \leq n^{\alpha+o(1)}$, and thus $\log(\mathbb{E}(e(\alpha)))/\log n \leq \alpha + o(1)$.

For the lower bound, note that all volumes of influence up to time $T = (A_1/2)n^\alpha$ have (deterministically) volume at least $2n^\alpha$. Therefore, a positive fraction of all edges generated until time T are long, and so *a.a.s.* $\Omega(n^\alpha)$ is a lower bound for the number of long edges and the theorem is finished. \square

REFERENCES

- [1] W. Aiello, A. Bonato, C. Cooper, J. Janssen, and P. Prałat, A spatial web graph model with local influence regions, *Internet Mathematics* 5 (2009), 175–196.
- [2] A. Barabási and R. Albert, Emergence of scaling in random networks, *Science* 286 (1999), 509–512.
- [3] J. Bichteler and E. Eaton, The combined use of bibliographic coupling and cocitation for document retrieval, *JASIST* 31(4) (1980), 278–284.
- [4] A. Bonato, J. Janssen, and P. Prałat, Geometric protean graphs, *Internet Mathematics* 8 (2012), 2–28.
- [5] A. Bonato, J. Janssen, and P. Prałat, The geometric protean model for on-line social networks, In: *Proc. 8th Workshop on Algorithms and models of the Web graph*, (WAW 2010), R. Kumar, R. Sivakumar (eds.), Springer LNCS 6516, 2010, pp. 110–121.
- [6] M. Bradonjic, A. Hagberg, and A.G. Percus, The structure of geographical threshold graphs, *Internet Mathematics* 4 (2009), 113–139.
- [7] C. Cooper, A. Frieze, and P. Prałat, Some typical properties of the Spatial Preferred Attachment model, In: *Proc. 9th Workshop on Algorithms and Models for the Web Graph* (WAW 2012), Springer LNCS 7323, 2012, pp. 29–40.
- [8] J. Dean and M.R. Henzinger, Finding related pages in the World Wide Web, *Computer networks*, 31(11–16):1467–1479, 1999.

- [9] L. Ferretti and M. Cortelezzi, Preferential attachment in growing spatial networks, *Phys. Rev. E*, 84, (2011), 1:016103.
- [10] A. Flaxman, A.M. Frieze, and J. Vera, A geometric preferential attachment model of networks, *Internet Mathematics*, 3(2), (2006), 187–206.
- [11] A. Flaxman, A.M. Frieze, and J. Vera, A geometric preferential attachment model of networks II, *Internet Mathematics*, 4(1) (2008), 87–111
- [12] D.J. Higham, M. Rasajski, and N. Przulj, Fitting a geometric graph to a protein-protein interaction network, *Bioinformatics*, 24(8) (2008), 1093–1099.
- [13] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, Wiley, New York, 2000.
- [14] 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.
- [15] M.M. Kessler, Bibliographic coupling between scientific papers, *Am. Doc.*, 14 (1963) 10–25.
- [16] J. Kleinberg, Navigation in a small world, *Nature* (2000) 406:845.
- [17] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguńá, Hyperbolic Geometry of Complex Networks, *Phys. Rev. E* 82 (2010) 3:036106.
- [18] K.-K. Lai and Sh.-J. Wu, Using the patent co-citation approach to establish a new patent classification system, *Inf. Proc. Mgt.*, 41(2) (2005), 313–330.
- [19] N. Masuda, M. Miwa, and N. Konno, Geographical threshold graphs with small-world and scale-free properties, *Phys. Rev. E*, 71 (2005) 3:036108.
- [20] F. Menczer, Lexical and semantic clustering by Web links, *JASIST*, 55(14) (2004), 1261–1269.
- [21] G. Oliva and S. Panzneri, Modeling Real Networks with Deterministic Preferential Attachment, in: *Proc. of 19th Mediterranean Conference on Control and Automation (MED)* (2011) pp. 13–18.
- [22] H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents, *JASIST* 24(4) (1973), 265–269.
- [23] H. van den Esker. A geometric preferential attachment model with fitness, *arXiv*0801.1612v1 (2008).
- [24] R. Wilson, Properties of the Spatial Preferential Attachment Model, MSc thesis, Dalhousie University, Halifax, Canada, April 2009.

DEPARTMENT OF MATHEMATICS AND STATISTICS, DALHOUSIE UNIVERSITY, PO BOX 15000,
HALIFAX, NS, CANADA, B3H 4R2

E-mail address: janssen@mathstat.dal.ca

DEPARTMENT OF MATHEMATICS, RYERSON UNIVERSITY, TORONTO, ON, CANADA, M5B
2K3

E-mail address: pralat@ryerson.ca

DEPARTMENT OF MATHEMATICS AND STATISTICS, DALHOUSIE UNIVERSITY, PO BOX 15000,
HALIFAX, NS, CANADA, B3H 4R2

E-mail address: rwilson@mathstat.dal.ca