Asymmetric Distribution of Nodes in the Spatial Preferred Attachment Model

Jeannette Janssen[1], Paweł Prałat[2], and Rory Wilson[1]

[1] Dalhousie University, Halifax, Canada
{jeannette.janssen,rory.ross.wilson}@dal.ca
[2] Ryerson University, Toronto, Canada
pralat@ryerson.ca

Abstract. In this paper, a spatial preferential attachment model for complex networks in which there is non-uniform distribution of the nodes in the metric space is studied. In this model, the metric layout represents hidden information about the similarity and community structure of the nodes. It is found that, for density functions that are locally constant, the graph properties can be well approximated by considering the graph as a union of graphs from uniform density spatial models corresponding to the regions of different densities. Moreover, methods from the uniform case can be used to extract information about the metric layout. Specifically, through link and co-citation analysis the density of a node's region can be estimated and the pairwise distances for certain nodes can be recovered with good accuracy.

Keywords: Spatial Random Graphs, Spatial Preferred Attachment Model, Preferential Attachment, Complex Networks, Web Graph, Co-citation, Common Neighbours

1 Introduction

There has been a great deal of recent interest in modelling complex networks, a result of the increasing connectedness of our world. The hyperlinked structure of the Web, citation patterns, friendship relationships, infectious disease spread, these are seemingly disparate collections of entities which have fundamentally very similar natures.

Many models of complex networks—such as copy models and preferential attachment models—have a common weakness: the 'uniformity' of the nodes; other than link structure there is no way to distinguish the nodes. One family of models which overcomes this deficiency is spatial (or geometric) models, wherein the nodes are embedded in a metric space. A node's position—especially in relation to the others—has real-world meaning: the character of the node is encoded in its location. Similar nodes are closer in the space than dissimilar nodes. This distance has many potential meanings: in communication networks, perhaps physical distance; in a friendship graph, an interest space; in the World Wide Web, a topic space. As an illustration, a node representing a webpage on pet food would be closer in the metric space to one on general pet care than to one on travel.

The Spatial Preferred Attachment Model [1], designed as a model for the World Wide Web, is one such spatial model. Indeed, as its name suggests, the SPA Model combines geometry and preferential attachment. Setting the SPA Model apart is the incorporation of 'spheres of influence' to accomplish preferential attachment: the greater the degree of the node, the larger its sphere of influence, and hence the higher the likelihood of the node gaining more neighbours. The SPA model produces scale-free networks, which exhibit many of the characteristics of real-life networks (see [1,4]). In [9], it was shown that the SPA model gave the best fit, in terms of graph structure, for a series of social networks derived from Facebook.

As the motivation behind spatial models is the 'second layer of meaning'—the character of the nodes as represented by their positions in the metric space—we hope to uncover this layer through examination of the link structure. In particular, estimating the distance between nodes in the metric space forms the basis for two important link mining tasks: finding entities that are similar—represented by nodes that are close together in the metric space—and finding communities—represented by spatial clusters of nodes in the metric space. We show how a theoretical analysis of a spatial model can lead to reliable tools to extract the 'second layer of meaning'.

The majority of the spatial models to this point have used uniform random distribution of nodes in the space. However, considering the real-world networks these models represent, this concept is impractical: indeed, on a basic level, if the metric space represents actual physical space, and the nodes people, then we note that people cluster in cities and towns, rather than being uniformly spread across the land. More abstractly, there are more webpages on a popular topic, corresponding to a small area of our metric space, than for a more obscure topic. The development of spatial network models naturally then begins to incorporate varying densities of node distribution: both 'clumps' of higher/lower density, as well as gradually changing densities, are both possibilities.

Of the more important goals is that of community recognition: the discovery and quantification of characteristically (semantically) similar nodes.

In this work we generalize the SPA model to non-homogeneous distribution of nodes within the space. We assume very distinct regions of different densities, 'clusters'. We find they behave almost as independent SPA Models of parameters derived from the densities. Many earlier results from the SPA Model then translate easily to this asymmetric version and we begin the process of uncovering the geometry using link analysis.

1.1 Background and Related Work

Efforts to extract node information through link analysis began with a heuristic quantification of entity similarity: numerical values, obtained from the graph structure, indicating the relatedness of two nodes. Early simple measures of entity similarity, such as the Jaccard coefficient [12], gave way to iterative graph theoretic measures, in which two objects are similar if they are related to similar objects, such as SimRank [10]. Many such measures also incorporate co-citation, the number of common neighbours of two nodes, as proposed in the paper by Small [13].

The development of graph models, in particular spatial models—as explored in [3] using thresholds, in combination with protean graphs [2] and with preferential attachment [5, 7]—added another dimension to node information extraction. For example, in [6], the authors make inferences on the social space for nodes in a social network, using Bayesian methods and maximum likelihood. But in particular, the authors' previous paper, [8], used common neighbours in a spatial model of the World Wide Web [1] to explore the underlying geometry and quantify node similarity based on distance in the space. In this paper, we draw heavily from [4], which includes further results on the SPA model, and in particular from [8] and extend its results to a generalization that allows us to overcome the reliance on uniform random distribution of nodes in the space. Non-uniform distributions have also been explored in [11, 14], as we move to more realistic models.

1.2 The Asymmetric SPA Model

We begin with a brief description of our Asymmetric SPA model. The model presented here is a generalization of the SPA model introduced in [1], the main difference being that we allow for an inhomogeneous distribution of nodes in the space.

Let S be the unit hypercube in $\mathbb{R}^m$, equipped with the torus metric derived from the Euclidean norm, or any equivalent metric. The nodes $\{v_t\}_{t=1}^n$ of the graphs produced by the SPA model are points in S chosen via an m-dimensional point process. Most generally, the process is given by a probability density function ρ; ρ is a measurable function such that $\int_S \rho d\mu = 1$. Precisely, for any measurable set $A \subseteq S$ and any t such that $1 \leq t \leq n$, $\mathbb{P}(v_t \in A) = \int_A \rho d\mu$.

In fact, we will restrict ourselves to probability functions that are *locally constant*. Precisely, we assume that the space $S = [0,1)^m$ is divided into k^m equal sized hypercubes, where k is a constant natural number. Each hypercube is of the form $I_{j_1} \times I_{j_2} \times \cdots \times I_{j_m}$ $(0 \leq j_1, j_2, \ldots, j_m < k)$, where $I_j = [j/k, (j+1)/k)$. Note that any density function ρ can be approximated by such a locally constant function, so that this restriction is justified.

To keep notation as simple as possible, we assume that each hypercube is labelled $\mathcal{R}_\ell$, $1 \leq \ell \leq k^m$. Let ρ_ℓ be the density of $\mathcal{R}_\ell$, so the density function has value ρ_ℓ on $\mathcal{R}_\ell$. For any node v, let $\mathcal{R}(v)$ be the hypercube containing v, and let $\rho(v)$ be the density of $\mathcal{R}(v)$. Clearly, every hypercube has volume k^{-m}. Then the probability that a node v_t, introduced at time t, falls in $\mathcal{R}_\ell$ equals $q_\ell = \rho_\ell k^{-m}$, and the expected number of points in $\mathcal{R}_\ell$ equals $\rho_\ell k^{-m} n$. It is easy to see that $\sum_\ell q_\ell = 1$. Thus we model the point process as follows: at each time step t, one of the regions is chosen as the destination of v_t; region $\mathcal{R}_\ell$ is chosen with probability q_ℓ. Then, a location for v_t is chosen uniformly at random from $\mathcal{R}_\ell$.

The SPA model generates stochastic sequences for graphs $(G_t : t \geq 0)$ with edge set E_t and node set $V_t \subseteq S$. The in-degree of a node v at time t is given by $\deg^-(v,t)$. Likewise the out-degree is given by $\deg^+(v,t)$. The sphere of influence of a node v at time t is defined as the ball, centred at v, with total volume

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

where $A_1, A_2 > 0$ are given parameters. If $(A_1 \deg^-(v,t) + A_2)/t \geq 1$, then $S(v,t) = S$ and so $|S(v,t)| = 1$. We impose the additional restriction that $pA_1 \max_j \rho_j < 1$; this avoids regions becoming too dense. This property will be always assumed. The generation of a SPA model graph begins at time $t = 0$ with G_0 being the null graph. At each time step $t \geq 1$ (defined to be the transition from G_{t-1} to G_t), a node v_t is chosen from S according to the given spatial distribution, and added to V_{t-1} to form V_t. Next, independently, for each node $u \in V_{t-1}$ such that $v_t \in S(u, t-1)$, a directed link (v_t, u) is created with probability p, $p \in (0,1)$ being another parameter of the model.

Let $\delta(v)$ be the distance from v to the boundary of $\mathcal{R}(v)$. Let $r(v,t)$ be the radius of the sphere of influence of node v at time t. So if $r(v,t) \leq \delta(v)$, then $S(v,t)$ is completely contained in $\mathcal{R}(v)$ at time t. We see that

$$r(v,t) = (|S(v,t)|/c_m)^{1/m} = \left(\frac{A_1 \deg^-(v,t) + A_2}{c_m t}\right)^{1/m},$$

where c_m is the volume of the unit ball; for example, in 2-dimensions with the Euclidean metric, $c_2 = \pi$.

Our goal is to investigate typical properties of a graph G_n on n nodes, and to use these to infer the spatial layout of the nodes. As typical in random graph theory, we shall consider only asymptotic properties of G_n as $n \to \infty$. We say that an event in a probability space holds *asymptotically almost surely* (a.a.s.) if its probability tends to one as n goes to infinity.

2 Graph properties of the SPA model.

In the Asymmetric SPA model with a locally constant density function, the probability of an edge forming from a new node v_t to an existing node v at time t equals

$$\mathbb{P}((v_t, v) \in E(G_n)) = p \int_{S(v,t)} \rho d\mu = p \sum_{\ell} \rho_\ell \, |S(v,t) \cap \mathcal{R}_\ell|.$$

Thus, the stochastic process of edge formation in the Asymmetric SPA model is bounded below by the process in which the edge probability is governed by $p\rho_{\min}$, and bounded above by that with $p\rho_{\max}$, where $\rho_{\min}$ and $\rho_{\max}$ are, respectively, the smallest and the largest densities occurring. The bounds on the link probability $\mathbb{P}((v_t, v) \in E(G_n))$ lead to bounds on the expected value of the degree.

Theorem 1 *Let $\omega = \omega(n)$ be any function tending to infinity together with n. The expected in-degree at time t of a node v_i born at time $i \geq \omega$ is given by*

$$(1+o(1))\frac{A_2}{A_1}\left(\frac{t}{i}\right)^{p\rho_{\min}A_1}-\frac{A_2}{A_1} \leq \mathbb{E}(\deg^-(v_i,t)) \leq (1+o(1))\frac{A_2}{A_1}\left(\frac{t}{i}\right)^{p\rho_{\max}A_1}-\frac{A_2}{A_1}.$$

In the analysis of the original SPA model, we find that nodes born quite early have their spheres of influence typically shrinking rapidly, and nodes born late start with small spheres of influence. A node would have to be quite close to the boundary of its region with another for the effect of any other region to be felt. It seems reasonable to expect that the graph formed by nodes in a region $\mathcal{R}_\ell$ with local density ρ_ℓ behaves like an independent SPA model of density ρ_ℓ.

To be specific, assume that nodes in the SPA model do not arrive at fixed time instances t, but instead arrive according to a homogeneous Poisson process with rate 1. (This will not significantly change the analysis.) Then, the process inside a region $\mathcal{R}$ with density ρ will behave like a SPA model with the same parameters A_1, A_2 and p, but with points arriving according to a Poisson process with rate ρ. This means that in each time interval we expect ρ points to arrive, and the expected time interval between arrivals equals $1/\rho$. If we use v_t to denote the t-th node arriving, then the arrival time $a(t)$ of v_t is approximately t/ρ, and thus the volume of the sphere of influence of an existing node v at the time that v_t is born equals

$$|S(v,a(t))| = \frac{A_1\deg^-(v,a(t))+A_2}{a(t)} \approx \frac{\rho A_1\deg^-(v,a(t))+\rho A_2}{t}.$$

Thus, in the analysis of the degree of an individual node, we expect a node v in the asymmetric SPA model to behave like a node in the original SPA model with parameters $\rho(v)A_1$, $\rho(v)A_2$ instead of A_1, A_2, where the degree of node v at time t in the Asymmetric SPA model corresponds to the degree of a node at time $a(t)$ in the corresponding SPA model. The following theorems show that this is indeed the case.

Theorem 2 *Let $\omega = \omega(n)$ be any function tending to infinity together with n. The expected in-degree at time t of a node v_i born at time $i \geq \omega \log n$, with $\delta(v) \gg (\log n/i)^{1/m}$ is given by*

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

Theorem 3 *Let $\omega = \omega(n)$ be any function tending to infinity together with n, and let $\varepsilon > 0$. The following holds* a.a.s. *For every node v for which* $\deg^-(v,n) = k = k(n) \geq \omega \log n$ *and for which*

$$\delta(v) \geq (1+\varepsilon)\left(\frac{A_1k+A_2}{c_mn}\right)^{1/m},$$

it holds that for all values of t such that $\max\{t_v, T_v\} \leq t \leq n$,

$$\deg^-(v,t) = (1+o(1))k\left(\frac{t}{n}\right)^{p\rho(v)A_1}.$$

Times T_v and t_v are defined as follows:

$$T_v = n\left(\frac{\omega \log n}{k}\right)^{p\rho(v)A_1}, \quad t_v = (1+\varepsilon)\left(\frac{A_1 k}{\delta^m c_m n^{p\rho(v)A_1}}\right)^{\frac{1}{1-p\rho A_1}}.$$

The statement of the theorem is rather technical, so we lay it out conceptually:

- the condition on $\delta(v)$ ensures that at time n, $S(v,n)$ is completely contained in $\mathcal{R}(v)$ (the factor of $(1+\varepsilon)$ gives some extra room for argument),
- time T_v is the time node v has $\omega \log n$ neighbours, provided that the process behaves as we expect,
- time t_v is the time when the sphere of influence has shrunk to the point where it became completely contained in $\mathcal{R}(v)$, provided the process behaves well (again, with extra room due to the factor $(1+\varepsilon)$). This occurs at the moment when the expected radius of the sphere of influence is smaller than $\delta(v)$.

The implication of this theorem is that once a node accumulates $\omega \log n$ neighbours and its sphere of influence has shrunk so that it does not intersect neighbouring regions, its behaviour can be predicted with high probability until the end of the process, and is completely governed by its region, and no others.

We note that if $\max\{T_v, t_v\} = t_v$ for a node v, then at time T_v—the time v first reaches in-degree $\omega \log n$—its sphere of influence extends beyond the region of v. However, since *a.a.s.* no node has degree $\omega \log n$ at time $O(\omega \log n)$, it must be that $T_v \gg \omega \log n$. Thus at time T_v the radius of the sphere of influence of v is $O\left((\omega \log n / T_v)^{1/m}\right) = o(1)$. The implication is that, in order for $\max\{T_v, t_v\}$ to be equal to t_v, a node would have to be very close to the border, that is, $\delta(v) = o(1)$. So for most nodes under consideration, $\max\{T_v, t_v\} = T_v$, and they behave like in a uniform SPA model of density $\rho(v)$ as soon as their degree reaches $\omega \log n$. Further, of these nodes, those with $\deg(v,n) \geq \omega^2 \log n$ reach degree $\omega \log n$ at time $o(n)$, and so have $o(\deg(v,n))$ neighbours outside $\mathcal{R}(v)$.

We can use the results on the degree to show that each graph induced by one of the regions $\mathcal{R}_\ell$ has a power law degree distribution. Let $N_i(j,n)$ denote the number of nodes of degree j at time n in the region $\mathcal{R}_i$ and let $j_f = j_f(n) = \left(n/\log^8 n\right)^{\frac{p\rho_{\max} A_1}{4p\rho_{\max} A_1 + 2}}$.

Theorem 4 *A.a.s. the graph induced by the nodes in region $\mathcal{R}_\ell$ has a power law degree distribution with coefficient $1 + 1/p\rho_i A_1$. Precisely,* a.a.s. *for any $1 \leq i \leq k^m$ there exists a constant c_i such that for any $1 \ll j \leq j_f$,*

$$N_i(j,n) = (1+o(1))c_i j^{-(1+\frac{1}{p\rho_i A_1})} q_i n.$$

Moreover, a.a.s. *the entire graph generated by the Asymmetric SPA model has a degree distribution whose tail follows a power law with coefficient $1 + 1/p\rho_{\max} A_1$.*

The number of edges also validates our hypothesis that a region of a certain density behaves almost as a uniform SPA model with adjusted parameters. In the SPA model with parameters $\rho_\ell A_1$, $\rho_\ell A_2$ and p, the average out-degree is approximately $\frac{p\rho_\ell A_2}{1-p\rho_\ell A_1}$, as per [1, Theorem 1.3]. The following theorem shows that the subgraph induced by one of the regions has the equivalent expected number of edges, and most edges have both endpoints in the same region. Moreover, this result shows why we need the condition $p\rho_{\max}A_1 < 1$. In fact, if $p\rho_{\max}A_1 \geq 1$, then the number of edges will grow superlinearly.

Theorem 5 *For a region $\mathcal{R}_\ell$ of density ρ_ℓ,* a.a.s. $|V(G_n) \cap \mathcal{R}_\ell| = (1+o(1))q_\ell n$. *Moreover,*

$$\mathbb{E}(\{(u,v) \in E(G_n) \,|\, u,v \in \mathcal{R}_\ell\}|) = (1+o(1))\frac{p\rho_\ell A_2}{1-p\rho_\ell A_1} q_\ell n.$$

Furthermore, a.a.s.

$$|\{(u,v) \in E(G_n) : \mathcal{R}(u) \neq \mathcal{R}(v)\}| = o(n),$$

i.e. the number of edges that cross the boundary of $\mathcal{R}_\ell$ is of smaller order than the number of edges completely contained in the region.

Our ultimate goal is to derive the pairwise distances between the nodes in the metric space through an analysis of the graph. The following theorem, obtained using the approach of [8], provides an important tool. Namely, it links the number of common in-neighbours of a pair of nodes to their (metric) distance. Using this theorem, we can then infer the distance from the number of common in-neighbours. Let $cn(u,v)$ denote the number of common in-neighbours of two nodes u and v.

The theorem distinguishes three cases. If u and v are relatively far from each other, then *a.a.s.* they will have no common neighbours. If the nodes are very close, then the number of common neighbours is approximately equal to a fraction p of the degree of the node of smallest degree. The third case provides a 'sweet spot' where the number of common neighbours is a direct function of the metric distance and the degrees of the nodes. For any two nodes u and v, let $cn(u,v,t)$ denote the number of common in-neighbours of u and v at time t.

Theorem 6 *Let $\omega = \omega(n)$ be any function tending to infinity together with n, and let $\varepsilon > 0$. The following holds* a.a.s. *Let u and v be nodes of final degrees* $\deg(u,n) = k$ *and* $\deg(v,n) = j$ *such that $\mathcal{R}(u) = \mathcal{R}(v)$, and $k \geq j \geq \omega^2 \log n$.*

Let $\rho = \rho(v)$ and let $T_v = n\left(\omega \log n / j\right)^{p\rho A_1}$, and assume that

$$\delta(v)^m \geq cj \text{ and } \delta(u)^m \geq ck, \text{ where } c = (1+\varepsilon)\left(\frac{A_1}{c_m n^{p\rho A_1} T_v^{1-p\rho A_1}}\right).$$

Let $d(u,v)$ be the distance between u and v in the metric space. Then, we have the following result about the number of common in-neighbours of u and v:

Case 1. If for some $\varepsilon > 0$

$$d(u,v) \geq \varepsilon \left(\frac{\omega \log n (k/j)}{T_v} \right)^{1/m}$$

then $cn(u,v,n) = O(\omega \log n)$.

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

$$d(u,v) \leq \left(\frac{A_1 k + A_2}{c_m n} \right)^{1/m} - \left(\frac{A_1 j + A_2}{c_m n} \right)^{1/m} = O\left(\left(\frac{k}{n} \right)^{1/m} \right),$$

then $cn(u,v,n) = (1+o(1))pj$. *If* $k = (1+o(1))j$ *and* $d(u,v)^m \ll (k/n) = (1+o(1))(j/n)$, *then* $cn(u,v,n) = (1+o(1))pj$ *as well.*

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

$$\left(\frac{A_1 k + A_2}{c_m n} \right)^{1/m} - \left(\frac{A_1 j + A_2}{c_m n} \right)^{1/m} < d(u,v) \ll \left(\frac{\omega \log n (k/j)}{T_v} \right)^{1/m},$$

then

$$cn(u,v,n) = C i_k^{-\frac{(p\rho A_1)^2}{1-p\rho A_1}} i_j^{-p\rho A_1} d(u,v)^{-\frac{mp\rho A_1}{1-p\rho A_1}} \left(1 + O\left(\left(\frac{i_k}{i_j} \right)^{p\rho A_1/m} \right) \right), \tag{1}$$

where $i_k = n \left(\frac{A_1}{A_2} k \right)^{-\frac{1}{p\rho A_1}}$ *and* $i_j = n \left(\frac{A_1}{A_2} j \right)^{-\frac{1}{p\rho A_1}}$, *and* $C = pA_1^{-1} A_2^{\frac{1}{1-p\rho A_1}} c_m^{-\frac{p\rho A_1}{1-p\rho A_1}}$. *If* $k = (1+o(1))j$ *and* $\varepsilon (k/n)^{1/m} < d(u,v) \ll (\omega \log n / T_v)^{1/m}$ *for some* $\varepsilon > 0$, *then*

$$cn(u,v,n) = \Theta \left(i_k^{-\frac{(p\rho A_1)^2}{1-p\rho A_1}} i_j^{-p\rho A_1} d(u,v)^{-\frac{mp\rho A_1}{1-p\rho A_1}} \right).$$

3 Reconstruction of Geometry

We set out to discover the character of nodes in a network purely through link structure, and to quantify the similarities. Spatial models allow us a convenient definition of similarity: distances between nodes. In examining the SPA model, the number of common neighbours allows us to uncover pairwise distances, a first step in the reconstruction of the geometry.

Description of Model Used For simulations, we use an Asymmetric SPA model we call the *diagonal layout*, which has 4 'clusters' of identical high density, with $m = 2$. In the diagonal layout, $k = 4$ and the 4 regions (x,x), $1 \leq x \leq 4$, are dense, with the others sparse. We will use 'dense region' and 'sparse region' to denote the union of all regions with densities ρ_d and ρ_s, respectively. For ease of notation, we note that $\rho_s = 4/3 - \rho_d/3$ so it is enough to provide the value of ρ_d only. In Figure 1 we see an example of the diagonal layout with nodes and

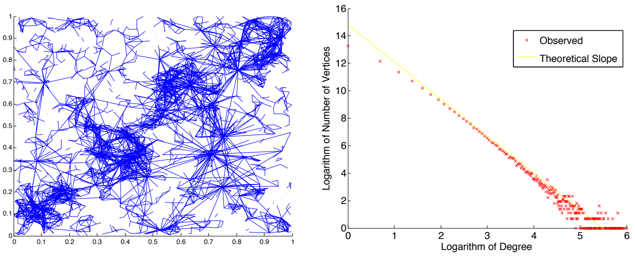

Fig. 1. Left: diagonal layout, $n = 1,000$, $p = 0.6$, $\rho_d = 1.6$, $A_1 = 0.7$, $A_2 = 2.0$; Right: degree distribution $n = 1,000,000$, $p = 0.7$, $\rho_d = 1.2$, $A_1 = 0.7$, $A_2 = 1.0$

edges, and we also see evidence that the densest region does dominate the power law degree distribution.

First we assume uniform density and apply the original estimator (Equation 7 from [8]) to our diagonal layout; the results are shown in the left in Figure 2. We eliminate those pairs we assume are in Case 1 (too close) and those in Case 2 (too far), by limiting our pairs to those with more than 10 common neighbours and fewer than $p/2\deg(v, n)$. This leaves 2270 pairs. The figure shows that the approach fails, and that it leads to a consistent overestimate of the distance for the nodes. This is somewhat counterintuitive, but the trouble lies with the estimator for a node's age based on in-degree: a node in $\mathcal{R}_d$ is thought to be much older than it actually is and confounds the distance estimator.

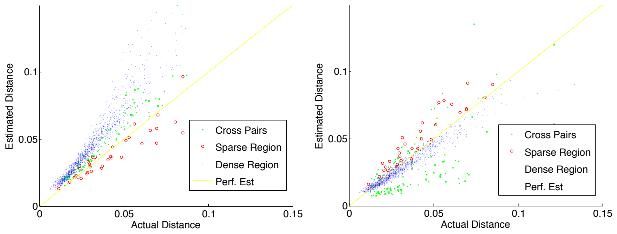

Fig. 2. SPA model, $n = 100,000$, diagonal layout, $p = 0.7$, $\rho_d = 1.6$, $A_1 = 0.7$, $A_2 = 2.0$, actual vs. estimated distances for pairs of nodes; Left: using original estimator; Right: using new estimator, density known

More precision is needed to take into account the varying densities. Examining Theorem 6, we note that for Case 3, equation (1) can be used to obtain an estimate $\hat{d}$ of the distance between a pair of nodes. For a pair of nodes u, v which are both in a region of density ρ, and their distance is such that Case 3 applies, this estimate is given by:

$$\hat{d}(u,v) = C_1(cn(u,v))^{-\frac{(1-p\rho A_1)}{mp\rho A_1}} k^{1/m} j^{\frac{(1-p\rho A_1)}{p\rho A_1 m}} \tag{2}$$

where $C_1 = (nc_m)^{-1/m} p^{-\frac{(1-p\rho(v_{(j)})A_1)}{mp\rho(v_{(j)})A_1}} A_1^{1/m} A_2^{\overline{\frac{2}{mp\rho(v_{(j)})A_1}}}$ and $k = \deg(u,n)$ and $j = \deg(v,n)$, with $k \geq j$.

Using the same simulation results, we compare the estimated distance using Equation 2 vs. actual node distance. Note we use our calculated density for each node to determine their estimated ages, but use the calculated density of the node of higher degree in the distance formula. The results seen on the right in Figure 2 indicate that our new estimator is quite accurate in predicting distances for some pairs of nodes, given all the parameters of the model, except for the cross-border pairs.

3.1 Estimating the density

In real-world situations, we cannot assume to know the density of the region containing a given node. In fact, the density of the region containing a node is an important part of the 'second layer of meaning' which we aim to extract from the graph. Therefore, in order to use our estimator for the distances between the nodes, we need to be able to use the graph structure to estimate the densities.

Using the theoretical results obtained from the previous section, we see that we can use the out-degrees of the in-neighbours of v to estimate the density of $\mathcal{R}(v)$. As per Theorem 5, the average out-degree in $\mathcal{R}_\ell$ is approximately $\frac{p\rho_\ell A_2}{1-p\rho_\ell A_1}$. Simulations confirm this expected value. Running sets of parameters 10 times each, we observe that if $p\rho_{\max}A_1 \leq 0.75$, the number of edges per region are within 90% of the expected value, on average. For $0.75 < p\rho_{\max}A_1 \leq 0.8$, the number of edges is within 75% of expected. For $p\rho_{\max}A_1 > 0.8$ we start to see deviation, as our expression for the expected number of edge in the densest region becomes 'unbounded', i.e. the denominator starts to approach 0. The number of edges that cross the border from sparse to dense, or between clusters, is consistently seen to be much smaller in order than the edges within each region.

Thus, if we have a large enough set of nodes from the same region, then we can use the formula above to estimate the density of the region. Consider a node v, and make two assumptions: (i) almost all neighbours of v are contained in $\mathcal{R}(v)$, and (ii) the neighbours of v form a representative sample of all nodes of $\mathcal{R}(v)$. Simulations show that these assumptions are justified and allow us to make an estimate for $\rho(v)$.

Set $\overline{\deg}^+(N^-(v))$ to be the average out-degree of the in-neighbours of v. Assuming the in-neighbours of v are also in $\mathcal{R}(v)$ (a fair assumption, given our earlier theorems), an estimator for the density can be derived from the average out-degree:

$$\hat{\rho(v)} = \frac{\overline{\deg}^+(N^-(v))}{pA_2 + pA_1\overline{\deg}^+(N^-(v))}.$$

We see in the histograms in Figure 3 (Left and Center) that the average out-degree of a node's in-neighbours in the dense region of the diagonal layout is quite accurate, but for the sparse region, the average out-degree is higher than expected. Displayed are the results for nodes with $\deg^{-}(v) \geq 10$. The calculated theoretical value for the average out-degree of in-neighbours for a node in the dense region is 5.85, and in the sparse, 1.45. This translates in $\rho_d = 1.6$ and $\rho_s = 0.8$. We see peaks that are quite accurate for the dense region, but translated to the right for the sparse region. Likely, those are sparse region nodes located close to the border; our condition on the minimum degree favours the 'rich' sparse region nodes.

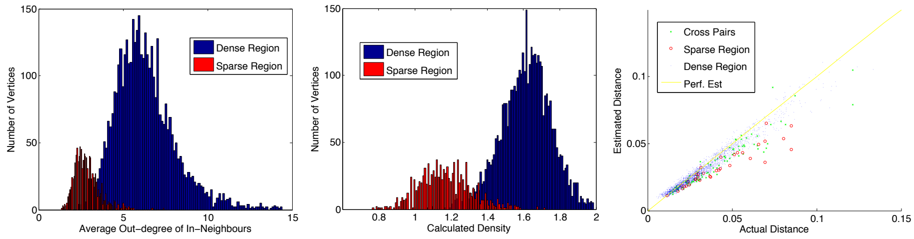

Fig. 3. Diagonal layout, $n = 100,000$, $\rho_d = 1.6$, $A_1 = 0.7$, $A_2 = 2.0$; Left: $p = 0.6$, average out-degree of the in-neighbours; Center: $p = 0.6$, calculated density from average out-degree; Right: $p = 0.7$, using estimated density from the node of greater final degree, all other parameters known

Finally, we use $\hat{\rho}$, and knowing all other parameters, to calculate the distance between the nodes based on number of common neighbours, Equation 2, using the same simulation results as earlier. Again note we use our calculated density for each node to determine their estimated ages, but use the calculated density of the node of higher degree in the distance formula. Using the lower degree node gives similar results. The results are seen in Figure 3 (Right). We obtain very good agreement between calculated and estimated densities.

4 Conclusion

Our analysis of a SPA model with non-uniform random distribution of nodes reveals almost independent clusters of nodes. Expected degree, degree distribution and number of edges behave as they would in localized SPA models with 'adjusted' parameters. It is not examined here, but it is suspected that these adjusted parameters extend to other existing results on the SPA Model such as the small world property and spectral properties: this is a goal of future work. The main result of the paper is that, by using the average out-degree of the in-neighbours of a node, an estimate of its region's density can be obtained. With this density in hand, the examination of common neighbours for pairs of nodes allows us to find their distances in the metric space. Currently, the number of pairs of nodes for which we have distances is quite limited due to the nature of the spheres of influence: that they either start small or shrink rapidly results in only those pairs that are very close having a significant number of common neighbours. Attempts to increase our information could include the use of path lengths, second neighbourhoods, etc.

Although the theoretical results are interesting in and of themselves, further work can be done in examining their validity in the context of real networks, i.e. recovering *meaningful* distances for pairs of nodes. Early results using machine learning and graphlets show that the SPA Model can be an accurate representation of social networks [9]; it would be ideal to extend our knowledge of the accuracy of the SPA Model, in particular the Asymmetric SPA Model, to other complex networks. In the context of real networks we may be able to further examine potentially 'anomalous nodes, such as those with shifting positions, or those with dual identities.

Our ultimate goal is reverse engineering: given the link structure of a graph, and assuming it could be modelled by the SPA model, we would be able to completely reconstruct the underlying spatial reality, a method of profound application. For example, knowing the hyperlink structure of a part of the Web, and assuming that it is well represented by the SPA model, we will be able to use this information to create a topic map of the pages. We will have developed a very powerful tool for prediction in the Web, with both economic and sociological benefits, such as improved web search and the discovery of cyber-communities.

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(1-2):175–196, 2008.
2. A. Bonato, J. Janssen, and P. Prałat. Geometric protean graphs. *Internet Mathematics*, 8(1-2):2–28, 2012.
3. M. Bradonjić, A. Hagberg, and A. Percus. Giant component and connectivity in geographical threshold graphs. In *Proceedings of the 5th Workshop on Algorithms and Models for the Web Graph*, WAW'07, pages 209–216. Springer, 2007.
4. C. Cooper, A. Frieze, and P. Prałat. Some typical properties of the spatial preferred attachment model. In *Proceedings of the 9th Workshop on Algorithms and Models for the Web Graph, WAW'12*, pages 29–40. Springer, 2012.
5. A. Flaxman, A. Frieze, and J. Vera. A geometric preferential attachment model of networks. In *Proceedings of the 3rd Workshop on Algorithms and Models for the Web Graph, WAW'04*, pages 44–55. Springer, 2004.
6. P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2001.
7. E. Jacob and P. Mörters. Spatial preferential attachment networks: Power laws and clustering coefficients. *ArXiv e-prints*, October 2012.
8. J. Janssen, P. Prałat, and R. Wilson. Geometric graph properties of the spatial preferred attachment model. *Advances in Applied Mathematics*, 50(2):243–267, 2013.

9. Jeannette Janssen, Matt Hurshman, and Nauzer Kalyaniwalla. Model selection for social networks using graphlets. *Internet Math*, 8(4):338–363, 2012.
10. G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *Knowledge Discovery and Data Mining*, pages 538–543, 2002.
11. J. Jordan. Geometric preferential attachment in non-uniform metric spaces. *ArXiv e-prints*, August 2012.
12. M. Kobayakawa, S. Kinjo, M. Hoshi, T. Ohmori, and A. Yamamoto. Fast computation of similarity based on jaccard coefficient for composition-based image retrieval. In *Proceedings of the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '09, pages 949–955. Springer-Verlag, 2009.
13. H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.
14. J. Zhang. Growing Random Geometric Graph Models of Super-linear Scaling Law. *ArXiv e-prints*, December 2012.