Estimating node similarity from co-citation in a spatial graph model

Jeannette Janssen
Dalhousie University
Halifax, NS, Canada
janssen@mathstat.dal.ca

Paweł Prałat
West Virginia University
Morgantown, WV, USA
pralat@math.wvu.edu

Rory Wilson
Dalhousie University
Halifax, NS, Canada
wilson.rory@gmail.com

Keywords

Node similarity, co-citation,bibliographic coupling,link analysis, complex networks, spatial graph model, SPA model

ABSTRACT

Co-citation (number of nodes linking to both of a given pair of nodes) is often used heuristically to judge similarity between nodes in a complex network. We investigate the relation between node similarity and co-citation in the context of the Spatial Preferred Attachment (SPA) model. The SPA model is a spatial model, where nodes live in a metric space, and nodes that are close together in space are considered similar, and are more likely to link to one another.

Theoretical analysis of the SPA model leads to a measure to estimate spatial distance from the link information, based on co-citation as well as the degrees of both nodes. Simulation results show this measure to be highly accurate in predicting the actual spatial distance.

1. INTRODUCTION

Studies of self-organizing networks of various kinds — the World Wide Web, citation graphs, online social networks, biological networks — have given convincing evidence that a significant amount of information about the entities represented by the nodes can be derived from the link environment of those nodes. A central question in such studies is how to extract information about similarity between node-entities from the link structure. This *link similarity* can be used as a complementary indication of similarity between nodes 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). It can also be used as the basis to identify *communities*, or clusters, of similar nodes.

One of the earliest measures of link similarity, proposed by Small in 1973 in a paper in library science is *co-citation* [13]. The co-citation measure of a pair of scientific papers is defined by the number of times both papers are both cited by another paper. In the wider context of complex networks, the co-citation measure of a pair of nodes is given by the number of nodes that link to both nodes of the pair. In terms of graph theory, this measure is also called the number of *common (in)-neighbours* of a pair of nodes. Co-citation, and the related measure of bibliographic coupling (from [8]) 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, 4, 12, 10].

While it is intuitively clear that the number of common neighbours should give an indication of node similarity, it is not so obvious how precisely this number should be interpreted. It seems plausible that the co-citation measure should somehow be scaled by the degree of the nodes: the fact that a pair of nodes has ten common neighbours will have different importance depending on whether the number of neighbours of each node is in the tens, or in the thousands. This consideration is especially important, since studies show that most of the networks under consideration have a degree distribution with a *power law* tail, which means that extremely high degrees occur in the network, although they are less common. Moreover, there is no "typical" node degree.

This paper shows that it is possible to quantify the relationship between the co-citation measure and node similarity, if we assume a generative model for the network where the nodes are embedded in an underlying space that represents their similarity. Precisely, we assume a *spatial graph model*, where nodes are points in a metric space, and the links are generated through a stochastic process that is influenced by the relative position of the nodes in space. The metric space is meant to represent the underlying reality of the entities represented by the nodes. Thus, if the nodes represent scientific papers, the space could be interpreted as the word-document space or some derivative thereof. If the nodes are individuals in a social network, the space could represent physical space, or an artificial "interest space" that represents the interests and activities of each individual. In protein-protein interaction networks, the space may represent the chemical properties of the proteins. The key assumption is that similar nodes will be close together in space, and that such nodes are more likely to link to each other than nodes that are far apart. Thus, the generated graph encodes, to some extent, the relative positions of the nodes, and it should be possible to extract information about the node positions from the link structure of the graph.

SAC'10 March 22-26, 2010, Sierre, Switzerland.
Copyright 2010 ACM 978-1-60558-638-0/10/03 ...$10.00.

A number of spatial models have been proposed recently [1, 2, 5, 6, 11, 7]. In most spatial models, however, the relationship between spatial distance and link formation is determined by a simple threshold function: a link is possible if vertices are within a prescribed distance t of each other, and impossible otherwise. For such models, information from the graph will only be able to determine whether the nodes are within distance t or not. In [9], a model is given where nodes are embedded in hyperbolic space. The Internet is embedded in this space using maximum likelihood.

This paper is based on the Spatial Preferred Attachment (SPA) model from [1]. In the SPA model, link formation is determined by spatial distance, as well as by the number of in-links of the receiving nodes. Thus, if the receiving node is highly popular, long links may occur. However, short links will be more common. The SPA model can therefore be seen as a plausible model for graphs governed by the Preferential Attachment principle (high degree nodes are more attractive), and where the length of an edge depends on the popularity of the receiving node. An example is the Web graph: popular web pages tend to receive more links, even if they are not that closely related to the web page of the originator of the link. While the results in this paper apply only to graphs generated by the SPA model as described, we believe that the results apply in general to graphs governed by those two principles.

The SPA model will be precisely described in Section 2. A theoretical analysis of a modified version of the SPA model, presented in Section 3, leads to a formula to predict the spatial distance between nodes, based on the number of common neighbours. In Section 4, the validity of the estimator is tested on a large set of simulating data, generated using the SPA model. By comparing real vs. estimated distance we will show that our estimator gives highly dependable results.

2. THE SPA MODEL

The SPA model. proposed in [1], is a stochastic graph model for self-organizing complex networks with an underlying spatial reality. The model generates directed graphs according to the following principle. Nodes are points in a given metric space S. Each node v has a *sphere of influence*. A new node u can only link to an existing node 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 the size of the sphere of influence of a node is proportional to its in-degree.

In [1], the model is defined for a variety of metric spaces S. In this paper, we let S be the unit square in $\mathbb{R}^2$, equipped with the torus metric derived from the Euclidean metric. This means that for any two points x and y in S,

$$d(x,y) = \min\{||x-y+u||_2 \, : \, u \in \{-1,0,1\}^n\}.$$

The torus metric thus “wraps around” the boundaries of the unit square; this metric was chosen to eliminate boundary effects.

The parameters of the model consist of the *link probability* $p \in [0,1]$, and three positive constants A_1, A_2 and A_3, which, for technical reasons, must be chosen so that $pA_1 \leq 1$.

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 node v in G_t, and $\deg^+(v,t)$ its out-degree. We define the *sphere of influence* $S(v)$ of node v at time $t \geq 1$ to be the ball centered at v with volume $A(v,t)$ defined as follows:

$$A(v,t) = \frac{A_1 \deg^-(v,t) + A_2}{t + A_3}, \tag{1}$$

or $A(v,t) = 1$ if the right-hand-side of (1) is greater than 1. Note that, since we use the torus metric, this ball will always be contained in S.

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

It was shown in [1] that the SPA model produces graphs with a power law degree distribution, with exponent $1 + 1/pA_1$. From the analysis presented in the paper, we can deduce the expected in-degree of a node, as given in the following theorem.

THEOREM 2.1. *The expected in-degree at time t of a node born at time i, if $i \gg 1$ and $t \gg i$, as $t \to \infty$, is given by*

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

3. NUMBER OF COMMON NEIGHBOURS AND SPATIAL DISTANCE

The principles of the SPA model make it plausible that nodes that are close together in space will have a relatively high number of common neighbours. Namely, if two nodes are close together, their spheres of influence will overlap a great deal, and any new node falling in the intersection of both spheres has the potential to become a common neighbour. Thus, co-citation should indeed lead to a reliable measure of similarity, here represented by closeness in the metric space. In this section, we will quantify the relation between spatial distance and number of common in-neighbours.

The size of the sphere of influence of a node is a function of its in-degree, and is therefore a random variable. For our analysis, we will modify the model so that the size is instead a deterministic variable. The simulation results presented in the next section show that this simplification is justifiable.

Precisely, we assume that at each time t

$$A(v_i, t) = \frac{\left(\frac{t}{i}\right)^p}{t}. \tag{3}$$

Asymptotically, the right hand side of (3) corresponds to the expected size of the region of influence in the original SPA model, based on the expected in-degree as given in equation (2). For simplicity, we have set $A_1 = A_2 = A_3 = 1$; inclusion of these parameters would only alter the results by a multiplicative constant. With this assumption, the size of the sphere of influence of each node shrinks with each time step.

For the remainder of this section, the results apply to this modified version of the SPA Model.

The radius of the sphere of influence of node v_i at time t can now be deduced from (3). Since we are using the Euclidean torus metric, the sphere of influence is a ball with

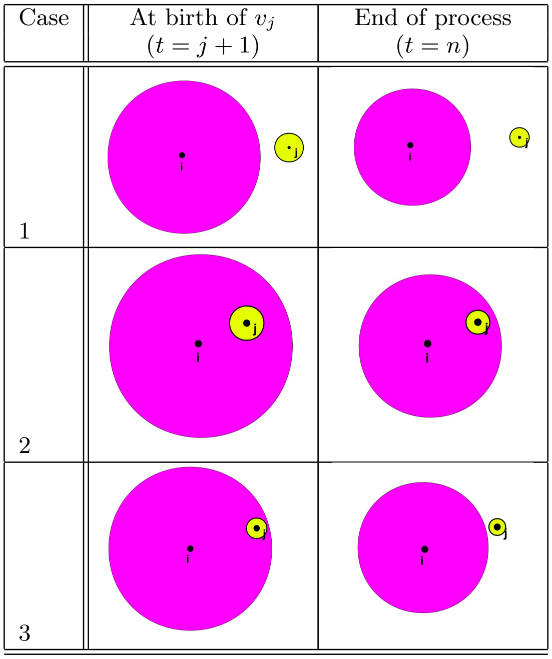

Figure 1: The three cases of Theorem 3.1

radius $r = r(v_i, t)$, satisfying $A(v_i, t) = \pi r^2$. Thus

$$r(v_i, t) = \sqrt{A(v_i, t)/\pi} = \pi^{-1/2} i^{-p/2} t^{-(1-p)/2}.$$

The term "common neighbour" here refers to common in-neighbours. Precisely, a node w is a common neighbour of nodes 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.

The following theorem gives bounds for the number of common in-neighbours, based on the spatial distance. There are three cases, depending on how the spheres of influence of v_i and v_j overlap, and when or whether they become disjoint. Figure 1 gives a pictorial representation of the three cases.

THEOREM 3.1. *Consider nodes v_i and v_j $(1 \leq i < j \leq n)$, in a graph generated by the SPA model as given, and let d be the distance between v_i and v_j according to the Euclidean torus metric. Then*

1. *If $d > r(v_i, j+1) + r(v_j, j+1)$, then v_i and v_j can have no common neighbours.*

2. *If $d \leq r(v_i, n) - r(v_j, n)$, then the expected number of common neighbours equals $(1 + o(1))p(n/j)^p$.*

3. *If $r(v_i, n) - r(v_j, n) < d \leq r(v_i, j+1) + r(v_j, j+1)$, then*

$$\mathbb{E}\, cn(v_i, v_j, n) = p\pi^{-\frac{p}{1-p}} \left(i^{-\frac{p^2}{1-p}}\right) \left(j^{-p}\right) \left(d^{-\frac{2p}{1-p}}\right) \left(1 + O\left(\left(\frac{i}{j}\right)^{p/2}\right)\right) \quad (4)$$

PROOF. Note that in the modified SPA model, the probability that v_j received a link at time t equals $pA(v_j, t) = p(i^{-p})(t^{-(1-p)})$. Therefore,

$$\mathbb{E}\deg^-(v_j, t) = \sum_{\tau=j+1}^{t} p j^{-p} \tau^{-(1-p)} = (1 + o(1)) \left(\frac{t}{j}\right)^p. \quad (5)$$

Case 1: If $d > r(v_i, j+1) + r(v_j, j+1)$ then the spheres of influence of v_i and v_j never intersect, so v_i and v_j can have no common neighbours.

Case 2: If $r(v_j, n) + d \leq r(v_i, n)$, then the sphere of influence of v_j is contained in the sphere of influence of v_i during the entire process. Any node v_k that links to v_j must fall inside the sphere of influence of v_i as well, and thus has a probability p of also linking to v_i. Thus the expected number of common neighbours is $p\,\mathbb{E}\deg^-(v_j, n) = (1+o(1))p\left(\frac{n}{j}\right)^p$.

Case 3: If $r(v_i, n) - r(v_j, n) < d \leq r(v_i, j+1) + r(v_j, j+1)$, then the spheres of influence of v_j and v_j overlap when v_j is born, but at least part of $S(v_j)$ is outside $S(v_i)$ at time n.

Let t_1 be the first moment that $S(v_j)$ is not completely contained in $S(v_i)$, i.e. the smallest $t \geq j$ so that $d + r(v_j, t) > r(v_i, t)$. (Let $t_1 = j$ if $S(v_j)$ is not contained in $S(v_i)$ at the birth of v_j.) Let t_2 be the first moment that $S(v_j)$ and $S(v_i)$ are completely disjoint (or let $t_2 = n$ if $S(v_j)$ and $S(v_i)$ overlap at time n). Up to time t_1, each neighbour of v_j will also be a neighbour of v_i with probability p. From time t_2 to n, any neighbour of v_j cannot also be a neighbour of v_i. From time t_1 until time t_2, the probability that a neighbour of v_j becomes a neighbour of v_i is *at most* p.

Thus, $p\mathbb{E}\deg^-(v_j, t_1)$ and $p\mathbb{E}\deg^-(v_j, t_2)$ form a lower and an upper bound, respectively, on the expected number of common neighbours of v_i and v_j. It can be easily derived that t_1 and t_2 as defined above satisfy $t_1 = \max\{\lfloor \tau_1 \rfloor + 1, j\}$ and $t_2 = \min\{\lfloor \tau_2 \rfloor + 1, n\}$, where

$$\tau_1 = \pi^{-\frac{1}{1-p}} \left(i^{-\frac{p}{1-p}}\right) \left(d^{-\frac{2}{1-p}}\right) \left(1 - \left(\frac{i}{j}\right)^{p/2}\right)^{\frac{2}{1-p}}$$

$$\tau_2 = \pi^{-\frac{1}{1-p}} \left(i^{-\frac{p}{1-p}}\right) \left(d^{-\frac{2}{1-p}}\right) \left(1 + \left(\frac{i}{j}\right)^{p/2}\right)^{\frac{2}{1-p}}$$

Combined with equation (5) about the expected degree, this leads to the the following bounds, which hold within a $(1 + o(1))$ term:

$$\pi^{-\frac{p}{1-p}} \left(i^{-\frac{p^2}{1-p}}\right) \left(j^{-p}\right) \left(d^{-\frac{2p}{1-p}}\right) \left(1 - \left(\frac{i}{j}\right)^{p/2}\right)^{\frac{2p}{1-p}}$$
$$\leq \mathbb{E}\, cn(v_i, v_j, n) \leq$$
$$\pi^{-\frac{p}{1-p}} \left(i^{-\frac{p^2}{1-p}}\right) \left(j^{-p}\right) \left(d^{-\frac{2p}{1-p}}\right) \left(1 + \left(\frac{i}{j}\right)^{p/2}\right)^{\frac{2p}{1-p}}.$$

The result follows from the fact that

$$\left(1 \pm \left(\frac{i}{j}\right)^{p/2}\right)^{\frac{2p}{1-p}} = 1 + O\left(\left(\frac{i}{j}\right)^{p/2}\right).$$

□

For any pair of vertices v_i and v_j, the events that v_i and v_j receive a common neighbour in time step t are independent for all $t > j$. Standard results such as the Chernoff bound can therefore be use to show that the actual number of common neighbours is concentrated around the expected value given in Theorem 3.1, provided the value is sufficiently large.

4. ESTIMATING DISTANCE BASED ON NUMBER OF COMMON NEIGHBOURS

In this section, we test the predictive power of our theoretical results on data obtained from simulations. The data was obtained from a graph with 100K nodes. The graph was generated from points randomly distributed in the unit square in $\mathbb{R}^2$ according to the SPA model described in Section 1, with $n = 100,000$ and $p = 0.95$. It is important to note that the data was generated using the *original* SPA model as described in Section 2. Our computational results will show that the assumption that led to the simplified model was justified.

First of all, we show that a blind approach to using the cocitation measure does not work. From the description of the SPA model it is clear that there exists a correlation between the spatial distance and number of common in-neighbours of a given pair of nodes. However, as shown in Figure 2, when we plot spatial distance versus number of common neighbours without further processing, no relation between the two is apparent.

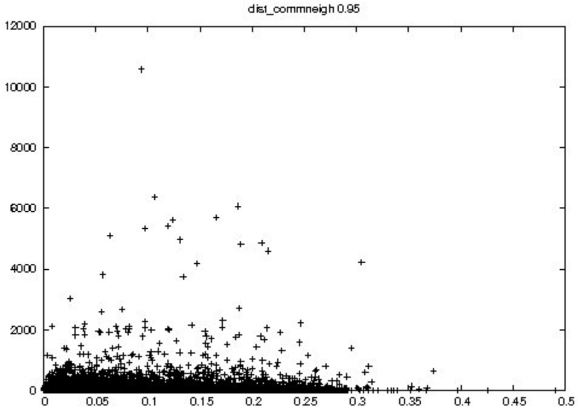

Figure 2: Actual distance vs. number of common neighbours.

The results from Theorem 3.1 lead to an estimate $\hat{d}$ of the spatial distance between two nodes, based on the number of common neighbours of the pair. (The spatial distance is the actual distance between the point in the metric space, i.e. the distance obtained from the Euclidean torus metric on the unit square.) Note that from case 1 and 2, we can only obtain a lower and upper bound on the distance, respectively. If two nodes v_i and v_j have no common neighbours, then we can assume we are in case 1, and thus $\hat{d} \geq r(v_i, j+1) + r(v_j, j+1)$. If $cn(v_i, v_j, n) \approx p \deg^-(v_j, n)$, then we are likely in case 2, and thus we get the upper bound $\hat{d} \leq r(v_i, n) - r(v_j, n)$. In order to eliminate case 1, we consider only pairs that have at least 20 common neighbours (19.2K pairs). To eliminate case 2, 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 2.4K pairs.

When we are likely in case 3, we can derive a precise estimate of the distance. We base our estimate on Equation (4), where we ignore the $O((\frac{j}{i})^{p/2})$ term. Namely, when i and j 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 $i \ll j$ the term is asymptotically negligible. The estimated distance between nodes v_i and v_j, given that their number of common neighbours equals k, is then given by

$$\hat{d} = \left(\pi^{-1/2} p^{\frac{1-p}{2p}}\right)\left(i^{-p/2}\right)\left(j^{-\frac{1-p}{2}}\right)\left(k^{-\frac{1-p}{2p}}\right).$$

Note that i and j appear in the formula above, so the estimated distance depends not only on the number of common neighbours of the two nodes, but also on their age. In our simulation data, the age of the nodes is known, and used in the estimate of $\hat{d}$. Figure 3 shows estimated distance vs. real distance between all pairs of nodes that are likely to be in case 3.

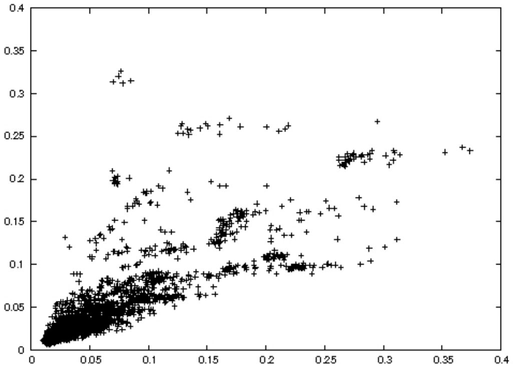

Figure 3: Actual distance vs. estimated distance for eligible pairs from simulated data, calculated using the age of both nodes.

While there is clearly some agreement between estimated and real distance, the variability in the results leads to the suspicion that the assumption made, namely that the sphere of influence is approximately equal to its expected value, may not be warranted. Indeed, further exploration of our simulation results revealed relatively large variability in the in-degree of a node of a given age. However, these deviations of individual nodes do not desturb the large scale pattern, since the degree distribution is entirely as expected. The solution suggested by these observations is to use, instead of actual age of a node, the estimated age based on its final in-degree. Thus, using the result from Theorem 2.1 for the SPA model, the birth time $\hat{a}(v)$ of a node v with in-degree k will be:

$$\hat{a}(v) = nk^{-1/p}.$$

Thus, we can compute $\hat{d}$ again, but this time based on the estimated birth times. This method has the added advantage that it can be more conveniently applied to real-life data, where the degree of a node is much easier to obtain than its age. Figure 4 again shows estimated vs. real distance for the exact same data set, but now estimated age is used in its calculation. This time, we see almost perfect agreement between estimate and reality.

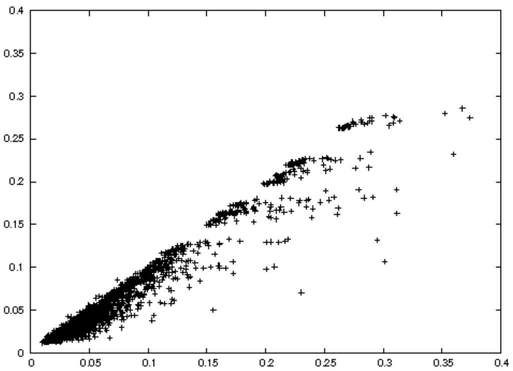

Figure 4: Actual distance vs. estimated distance for eligible pairs from simulated data, calculated using the in-degree of both nodes.

5. CONCLUSIONS AND FURTHER WORK

We have shown how a theoretical analysis of the SPA model leads to a highly accurate measure for the estimated spatial distance between nodes, based on the number of common neighbours. This shows how the assumption of a generative graph model can be used to obtain predictions about the underlying reality of the nodes from the link structure. While the results obtained apply only to a specific situation: a graph generated by the SPA model, and points uniformly distributed in the space, we believe strongly that the results, with some possible alterations, will be applicable to real life networks. Generally, we conjecture that the results apply to networks that satisfy the underlying principles of the SPA model: high degree nodes are more attractive, and links to popular nodes can span a longer distance in space. The next step will be to test our conjecture by applying the results obtained in this paper to a real life network which satisfies these general principles (for example, the citation graph, any social network, or part of the World Wide Web). Some of these graphs are readily available together with a number of text mining tools that can be used to measure how similar given two nodes are.

The methods presented in this paper are only valid for a subset of all pairs of nodes. In further work, we will investigate to what extent the distances between these pairs suffice to infer the relative placement of all nodes in the space. Specifically, it would be useful to know if, when the data is clustered, the distances available are sufficient to obtain the clusters. Also, further analysis may lead to alternative ways to estimate the distance for the pairs where our methods do not apply, based on, for example, graph distance or number of paths of given lengths between the nodes.

6. REFERENCES

[1] W. Aiello, A. Bonato, C. Cooper, J. Janssen, P. Prałat, A spatial web graph model with local influence regions, *Internet Mathematics* **5** (2009), 175–196.

[2] M. Bradonjic, A. Hagberg, A.G. Percus, Giant component and connectivity in geographical threshold graphs, *Proc. WAW 2007)*, LNCS **4863**, pp. 209–216.

[3] J. Bichteler, E. Eaton, The combined use of bibliographic coupling and cocitation for document retrieval, *JASIST* **31(4)** (1980), 278–284.

[4] J. Dean, M.R. Henzinger, Finding related pages in the World Wide Web, *Computer networks* **31(11–16)** (1999), pp. 1467–1479.

[5] A. Flaxman, A.M. Frieze, J. Vera, A geometric preferential attachment model of networks, *Internet Mathematics* **3(2)** (2006) pp. 187–206.

[6] A. Flaxman, A.M. Frieze, J. Vera, A geometric preferential attachment model of networks II, *Internet Mathematics* **4(1)** (2008), pp. 87–111.

[7] D.J. Higham, M. Rasajski, N. Przulj, Fitting a geometric graph to a protein-protein interaction network, *Bioinformatics* **24(8)** (2008), pp. 1093–1099.

[8] M.M. Kessler, Bibliographic coupling between scientific papers, *Am. Doc.* **14** (1963), pp. 10–25.

[9] D. Krioukov, F. Papadopoulos, A. Vahdat, M. Boguña, Curvature and Temperature of Complex Networks, *arXiv*0903.2584v2 (2009)

[10] K.-K. Lai, Sh.-J. Wu, Using the patent co-citation approach to establish a new patent classification system, *Inf. Proc. Mgt.* **41(2)** (2005), pp. 313–330.

[11] N. Masuda, M. Miwa, N. Konno, Geographical threshold graphs with small-world and scale-free properties, *Phys. Rev. E* **71(3)** (2005) 036108.

[12] F. Menczer, Lexical and semantic clustering by Web links, *JASIST* **55(14)** (2004), pp. 1261–1269.

[13] H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents, *JASIST* **24(4)** (1973), pp. 265–269.