Spatial models for virtual networks

Jeannette Janssen

Dalhousie University, Halifax, NS, B3H 3J5, Canada

Abstract. This paper discusses the use of spatial graph models for the analysis of networks that do not have a direct spatial reality, such as web graphs, on-line social networks, or citation graphs. In a spatial graph model, nodes are embedded in a metric space, and link formation depends on the relative position of nodes in the space. It is argued that spatial models form a good basis for link mining: assuming a spatial model, the link information can be used to infer the spatial position of the nodes, and this information can then be used for clustering and recognition of node similarity. This paper gives a survey of spatial graph models, and discusses their suitability for link mining.

1 Introduction

Through the advent of the Internet and especially the World Wide Web, huge repositories of data have become available in a naturally linked form. Examples are: on-line social networks, thematically coherent or domain-restricted segments of the World Wide Web, and electronic libraries of scientific papers. The links connecting entities in such data collections form a virtual network. The link structure of this network encodes information about the data represented by the nodes. *Link mining* is the process of extracting that information. Link mining can give information about the data collection when node-specific data is unavailable or private, as in on-line social networks, or can be combined with text mining in web graphs or citation networks to gain a better understanding of the data.

In order to interpret the structure of the network it helps to model the process that led to the formation of the network. The virtual networks that are the subject of this paper are *self-organizing*: they are not governed by central control or design, but formed by individual actions of autonomous agents: Facebook users, Web page designers, authors of scientific papers. Moreover, since the networks are virtual and link creation is free, there are no physical constraints that limit the link structure.

The first studies of the link structure of the Web revealed that virtual self-organizing networks exhibit a characteristic structure. The first models for such networks mainly aimed to generate graphs with a similar substructure. For a survey of such models, see [4, 3]. The principal properties observed were:

1. *A heavy tail degree distribution.* A characteristic of almost all virtual networks is that high degree nodes are relatively common. That is, the degree distribution $P(k)$, where $P(k)$ is the proportion of nodes of degree k, does

not fall off exponentially as k grows large. Often, the tail of the distribution follows a *power law*: $P(k) \sim k^{-\gamma}$ for some exponent γ which is usually between 2 and 3.

2. *The graphs are globally sparse, but locally dense.* The average degree is $O(\log n)$, where n is the size of the network, or can even be constant. On the other hand, locally the graph is denser than one would expect if the links were distributed randomly. This can be measured by the value of the *clustering coefficient*, which is the average density of the subgraphs induced by the neighbourhood of a node.
3. *Small distances between nodes.* The local density might lead one to expect that it can take a large number of hops to go from one node to the other, since many links remain local. However, this is not the case: the average distance between nodes is $O(\log n)$ or smaller.

While many of the graph models were succesful in reproducing some or all of the observed graph properties, they are based on an assumption that is incompatable with the concept of link mining. Generally, nodes in such models are indistinguishable. That is, the stochastic process that determines the link neighbourhood of a node is *only* based on the existing network, not on any individual property of the node. Link mining is based on the premise that nodes are *not* equal. The premise that it is possible to extract information about the nodes from the link structure implies that this information is present but hidden, and further, that this information influences the formation of the link structure. Thus, the link structure is a visible manifestation of an underlying, hidden reality.

Two principal tasks of link mining are those of detecting *similarity* between nodes, and of identifying *communities* of related nodes. Thus, the hidden reality should give information about node similarity. A natural way to model similarity is to assume that the nodes are embedded in a metric space, where the metric distance between nodes is a measure of the similarity between nodes. If two nodes are similar the metric distance between them will be small, and communities will correspond to spatial clusters.

Graph models where the nodes are embedded in a metric space and link formation is influenced by the metric distance between nodes are called *spatial models*. The main principle of spatial models is that nodes that are metrically close are more likely to link to each other. This is a formal expression of the intuitive notion we hold about virtual networks: Web links are likely to point to similar pages, people that share similar interests are more likely to become friends on Facebook, and a scientific paper mostly refers to papers on a similar topic. This paper gives an overview of spatial models and their suitability for link mining.

2 Spatial models with network-based link formation

In this section we review models where the link formation is directly influenced by the distance between the nodes. Precisely, let d be the metric of the space

in which the nodes are located. The probability that node v_i links to node v_j is a decreasing function of $d(v_i, v_j)$. A review of these models can also be found in Section 2.5 of [3]. The models have in common that link formation is *network based*, which means that the stochastic process according to wich the links attaching to a node are generated depends on the entire network, and the node can potentially link to any node in the network, albeit with increasingly smaller probability as the distance increases.

Most of the models presented in this Section were proposed as models for *spatial networks*, characterized in [3] as "networks whose nodes occupy a precise position in two-or three-dimensional Euclidean space, and whose edges are physical connections". Examples of such networks are: the Internet (where the nodes are the routers) and other physical communication networks, railway and road networks, electric power grids, and neural networks in the brain. In spatial networks, there are limitations on the network posed by the physical reality; in virtual networks such limitations are largely absent. Still, the principle that nodes that are close in space are more likely to link to each other holds both both types. We will argue that some of these spatial models fit the reality of virtual models.

Unless otherwise mentioned, we consider the space S in which the nodes are embedded to be the hypercube $[0,1]^D$, where D is the dimension. In order to eliminate boundary effects, we use the *torus metric* derived from any of the L_p norms. Formally, this means that for any two points x and y in S,

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

The torus metric thus "wraps around" the boundaries of the unit cube, so every point in S is equivalent.

In the models of spatial networks, the dimension D must equal 2 or 3, and the metric is the Euclidean metric (derived from the L_2 norm). For a realistic model of virtual networks we expect D to be higher, while the metrics derived from the L_∞ norm (determined by the largest coordinate) and the L_1 norm (determined by the sum of the coordinates) can be reasonable alternatives.

The nodes are embedded in S according to a given probability distribution. For ease of analysis, this distribution is often assumed to be uniform. However, a distribution that aims to model real data should contain clusters of closely spaced points.

The most straightforward spatial models are those where links between nodes are formed independently, and a link between two nodes that have distance r to each other is formed with probability $p(r)$. An early model for the Internet by Waxman [23] takes $p(r) = \beta exp(-r/\alpha c)$, where $\alpha, \beta \in (0,1)$ are parameters, and c is the maximum distance between any pair of points. A similar model in studied in [14]. The exponential decay in $p(r)$ implies that links substantially longer than αc are highly unlikely to occur. The graph distance between two nodes that are far apart in S dependes on the maximum metric distance that can be spanned by a link. A good model for virtual networks will need to include *long links*, i.e. links between nodes that are far apart in the space, in order to achieve the property of small average distance between nodes.

In order to generate networks that have the desired properties, especially a suitable degree distribution, the link neighbourhood should also depend on the link structure of the network. This can be done if the graph is generated node by node. Formally, starting from a small initial graph, at each time step t a new node v_t is generated and embedded in S according to the predetermined probability distribution. The new node is given a fixed number m of initial links. For each of the links, the probability $p(i,t)$ that an existing node v_i is chosen as the other endpoint is a function of the metric distance between v_i and v_t, and graph properties of v_i in the existing network; usually the node degree. The probability $p(i,t)$ is called the *link probability*.

In [26, 17, 20], the link probability is given as:

$$p(i,t) = \frac{deg(v_i)}{c(t)\, d(v_i, v_t)^\alpha},$$

where $\alpha > 0$ is a parameter, and $c(t)$ is a normalizing constant. In [26], this model is studied in one dimension ($D = 1$). Since we use the torus metric, we can imagine the nodes as being located in a circle. Thus, the model is a natural extension of the "small-world" network of Watts and Strogatz [22]. For this model, is determined exponentially that, for $\alpha < 1$ the degree distribution follows a power law, $P(k) \sim k^{-3}$, and for $\alpha > 1$ it is a stretched exponential, $P(k) \sim \exp(-bk^\gamma)$, where γ depends on α. For all values of α, the average distance between nodes is of order $\log n$.

In [17], the model is studied numerically in all dimensions, and a critical value α_c is determined so that for all $\alpha < \alpha_c$, the degree distribution follows a power law, while for $\alpha > \alpha_c$, it is a stretched exponential. It is also shown that the *link length distribution* $\ell(k)$, where $\ell(k)dk$ is the proportion of links that have length between k and $k + dk$, follows a power law.

These papers derived their results from a uniform distribution of points in space; in [20], a non-uniform distribution is studied. Here $D = 2$. Any new node is embedded randomly in S at distance r from the barycenter of the existing set of nodes, where r is chosen probabilistically according to a probability distribution $P(r) \sim r^{-(2+\beta)}$, where $\beta > 0$ is a parameter. Using methods from statistical mechanics, the authors numerically determine the values of the parameters for which the graph has a power law degree distribution.

In [2], the case where $p(i,t) \sim deg(v_i) \exp(-d(v_i, v_j)/r_c)$ is studied for $D = 2$. In this case, links significantly longer than r_c will be rare. In [24], a model for the Internet is proposed where $D = 2$, and

$$p(i,t) = \frac{deg(v_i)^\sigma}{c(t)\, d(v_i, v_t)^\alpha},$$

where both σ and α are parameters. Moreover, a non-uniform distribution of nodes in space is used which mimicks actual human population densities.

In [15], nodes are placed on a grid pattern in a 2-dimensional space. Each node is connected to its four nearest neighbours in the grid, and to one other node. The destination of the extra link is chosen from all other nodes with

probability $r^{-\alpha}$, where r is the distance between the origin and the destination node. Thus, this model can be considered as a grid graph superemposed on a spatial graph model with $p(i,t) = d(v_i, v_j)^{-\alpha}$, $D = 2$ and $m = 1$.

Most of the models just discussed can produce graph with the same properties as those observed in many virtual networks. However, their suitability for link mining, as discussed in the introduction, has not yet been investigated. Since link formation is directly dependent on the metric distance between nodes, it is reasonable to expect that precise inferences about the metric distances between nodes can be made from the network. For example, a set of nodes that are placed close together in the metric space will likely have higher than average link density.

Finally, we mention three spatial models that are based on different principles. In [25], nodes are linked independently with probability proportional to the dot product of the vectors reprenting the nodes. In [18], and later also [7], nodes are assigned random weights w_i, and two nodes v_i and v_j are linked precisely when a function of the weights and the distance, for example $(w_i + w_j)/d(v_i, v_j)$, exceeds a given threshold θ. In [8], each new node v_t links to the node v_i that minizes a function which is the convex combination of the graph distance of v_i to the center of the graph, and the metric distance between v_i and v_t.

3 Spatial models with node-based link formation

Virtual networks can be very large, so it is reasonable to assume that any user is only familiar with a small fraction of the network, and that this fraction consists of the nodes that are similar to the node associated with the user. In the context of spatial models, this implies that a new node can only see the part of the network that corresponds to nodes that are metrically close. The second class of models we survey are based on this principle.

The simplest model is the *random geometric graph* [19]. In this model, nodes are embedded in a metric space according to a given probability distribution, and two nodes are linked precisely when their distance is smaller than a given threshold value θ. The random geometric graph was proposed as a model for spatial networks in [11], for wireless multi-hop networks in [21], and for biological networks in [12]. In [16], an interesting variant of the model is presented, where the metric space is the hyperbolic space. In random geometric graphs, a link between two nodes gives exact binary information about whether the nodes are within distance θ of each other or not. Thus the graph distance should be highly correlated with the metric distance, up to multiples of θ. In [5], this relationship is confirmed for infinite random geometric graphs. In this paper also a variation of the model is proposed, where nodes that are within distance θ of each other are linked independently with probability p.

A variation of the random geometric graph model where the link probability is partially determined by the network was proposed by Flaxman *et al.* in [9]. Here, nodes join the network one by one, and each new node receives m links, and chooses its neighbours from among the nodes that are within distance θ

of it. The probability that a node v_i, which is within distance θ of the new node, receives a link, is proportional to its degree. As in the previous model, the threshold θ limits the length of a link, so that long links become impossible. In [10], the same authors extend the model so that, the hard threshold θ, is determined by a function which makes it less likely that a node receives a link if it is far removed from the new node.

The random geometric graph can also be represented as follows: each node is the centre of an associated sphere of radius θ. A new node can link to an existing node only when it falls within its associated sphere. The SPA model, proposed in [1], is based on the same principle, except that the radii of the spheres are not uniform, but depend on the in-degree of the node (the SPA model generates directed graphs). Precisely, each node v_i is the center of a sphere whose radius is chosen so that its volume equals

$$A(v_i, t) = \min\{\frac{A_1 \deg^-(v_i, t) + A_2}{t + A_3}, 1\}.$$

At each time step t, a new node v_t is created and embedded in S uniformly at random. For each node v_i so that v_t falls inside the sphere of v_i, independently, a link from v_t to v_i is created with probability p. Thus, a node with high degree is "visible" over a larger area of S than a node with small degree. It was shown in [1] that the SPA model produces graphs with a power law degree distribution, with exponent $1 + 1/pA_1$. Moreover, because of the unequal size of the spheres, long links can occur, but only to nodes of high degree. It is shown in the next section that the SPA model can be used to retrieve the metric distances between nodes from the network with high precision.

A related model, the geo-protean model, was proposed in [6]. Here, the size of the network is constant, but at each step a node is deleted and a new node is added. Nodes are ranked from 1 to n, and the sphere around a node is such that its volume is proportional to is rank raised to the power $-\alpha$. The link neighbourhood of the new nodes is determined in a similar way as in the SPA model. It is shown that the degree distribution, link density, average distance between nodes, and clustering behaviour are consistent with those observed in social networks.

4 Estimating distance from the number of common neighbours

In this section, we show how the metric distances between nodes can be estimated from their number of common neighbours. The model we use is the SPA model, described in the last section. The work presented here can be found in more detail in [13]. In this section, the dimension of S is assumed to be 2, and the parameters of the SPA model are $A_1 = A_2 = A_3 = 1$.

The SPA model produces directed graphs. 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. In

the SPA 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.

First of all, we show that a blind approach to using the co-citation 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 4, when we plot spatial distance versus number of common neighbours without further processing, no relation between the two is apparent.

The data presented in Figure 4 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, 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 the previous section (so the volume of the sphere of influence is proportional to the real degree, not the expected degree).

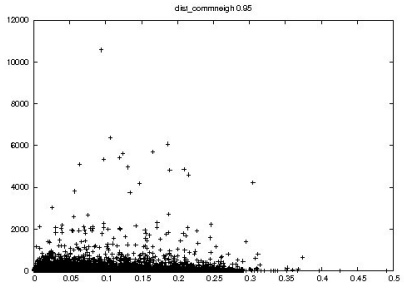

Fig. 1. Actual distance vs. number of common neighbours.

By analyzing the model, we can extract the relationship between the number of common neighbours and the metric distance. First, we make a simplifying assumption. In [1], it was shown that the expected degree, at time t, of node v_i born at time i is proportional to $(t/i)^p$. Assuming that the degree each node *equals* its expected degree, we obtain that the volume of the sphere of a node equals

$$A(v_i, t) = \frac{\left(\frac{t}{i}\right)^p}{t}. \tag{1}$$

The radius $r(v_i, t)$ of the sphere of influence of node v_i at time t can now be deduced from its volume as given above. Since we are using the Euclidean metric, this radius is given by:

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

The relationship between the number of common neighbours and the metric distance of two nodes v_i and v_j at distance $d(v_i, v_j) = d$ can now be given as follows:

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

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

These formulas lead to an estimate $\hat{d}$ of the metric distance between two nodes, based on the number of common neighbours of the pair. 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 our simulation, 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 (2), 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 4 shows estimated distance vs. real distance between all pairs of nodes that are likely to be in case 3.

While there is clearly some agreement between estimated and real distance, the results can be improved if we use, instead of the real age, the estimated age. The estimated time of birth $\hat{a}(v)$ of a node v which has in-degree k at time n 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.

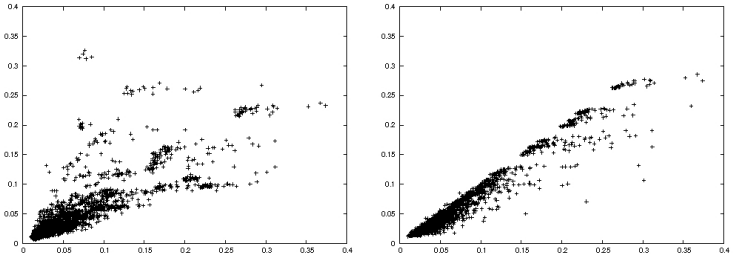

Fig. 2. Actual distance vs. estimated distance for eligible pairs from simulated data, calculated using the age (left) and estimated age from degree (right) of both nodes.

5 Conclusion

For the purpose of link mining, it is useful to assume a spatial model even when the data represented by the network have no direct spatial reality. The data can instead be considered to be embedded in a multi-dimensional space in such a way that metric closeness indicates similarity.

In this paper, I have given a survey of spatial models. Most of those models where conceived as models for networks that have a direct spatial reality. I propose that these models can be adapted to virtual networks, and that an analysis of these models can lead to precise instructions on how to infer information about the metric distances between nodes from the network data. As an example, I have shown in the last section how, in one model, the number of common neighbours gives a precise estimate of the metric distance.

Much further work in this field is needed, both theoretical and experimental. A theoretical analysis of the models in Section 2 may lead to useful measures for link mining. These measures should then be applied to real data. A challenge is to find data where a "ground truth" embedding of the nodes in metric space is present. One proposition is to use data from real spatial networks to validate the models and measures. Another suggestion is to work with citation graphs and Web graphs, and use the text information associated with the nodes (scientific papers or Web pages) to obtain an embedding of the nodes in space, for example through word-document representations and Latent Semantic Indexing. This embedding can then be compared with the embedding obtained through link mining.

References

1. W. Aiello, A. Bonato, C. Cooper, J. Janssen, P. Prałat. A spatial web graph model with local influence regions. *Internet Mathematics*, 5(1–2):175–196, 2009.
2. M. Barthelemy. Crossover from scale-free to spatial networks. *Europhysics Lett.* 63(6):915–921 (2003).
3. S. Boccaletti *et al.*: Complex networks: Structure and dynamics. *Physics Reports* 424:175–308, 2006.

4. A. Bonato: A Course on the Web Graph. American Mathematical Society, Providence, Rhode Island, 2008.
5. A. Bonato and J. Janssen. Infinite random geometric graphs. *Preprint* (2010).
6. A. Bonato, J. Janssen and P. Prałat. A geometric model for on-line social networks. *Preprint* (2010).
7. M. Bradonjic, A. Hagberg, A.G. Percus. The structure of geographical threshold graphs. *Internet Mathematics* 4(1–2):113–139, 2009.
8. A. Fabrikant, E. Koutsoupias and C.H. Papadimitriou. Heuristically Optimized Trade-offs: a new paradigm for power laws in the Internet. *In Proceddings of ICALP'02*, Springer LNCS 2380:110–122 (2002).
9. A. Flaxman, A.M. Frieze, J. Vera. A geometric preferential attachment model of networks. *Internet Mathematics*, 3(2):187–206, 2006.
10. A. Flaxman, A.M. Frieze, J. Vera. A geometric preferential attachment model of networks II. *Internet Mathematics*, 4(1):87–111, 2008.
11. C. Herrmann, M. Barthélemy and P. Provero. Connectivity distribution of spatial networks. *Phys. Rev.E* 68, 026128 (2003).
12. D.J. Higham, M. Rasajski, N. Przulj. Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics*, 24(8):1093–1099, 2008.
13. J. Janssen, P. Prałat, R. Wilson. Estimating node similarity from co-citation in a spatial graph model. In: *Proceedings of 2010 ACM Symposium on Applied Computing–Special Track on Self-organizing Complex Systems*, 5pp.
14. M. Kaiser and C.C. Hilgetag. Spatial growth of real-world networks. *Phys. Rev. E* 69, 036103 (2004).
15. J. Kleinberg. Navigation in a small world. *Nature* 406:845 (2000).
16. D. Krioukov, F. Papadopoulos, A. Vahdat, M. Boguña. Curvature and Temperature of Complex Networks. *arXiv*0903.2584v2, 2009.
17. S.S. Manna and P. Sen. Modulated scale-free network in Euclidean space. *Phys. Rev. E* 66, 066114 (2002).
18. N. Masuda, M. Miwa, N. Konno. Geographical threshold graphs with small-world and scale-free properties. *Phys. Rev. E*, 71(3):036108, 2005.
19. M. Penrose. Random Geometric Graphs. Oxford University Press, Oxford, 2003.
20. D.J.B. Soares, C. Tsallis, A.M. Mariz and L.R. da Silva. Preferential attachment growth model and nonextensive statistical mechanics. *Europhys. Lett.* 70(1), pp. 70–76 (2005).
21. X. Ta, G. Mao and B.D.O. Anderson. On the Properties of Giant Component in Wireless Multi-hop Networks. *In: 28th IEEE INFOCOM proceedings* (2009).
22. D.J. Watts and S.H. Strogatz. Collective dynamics of "small-world" networks. *Nature* 393(6684):409–410 (1998).
23. B.M. Waxman. Routing of multipoint connections. *IEEE J.Sel. Areas in Comm.*, 6(9):1617–1622, 1988.
24. , S.-H. Yook, H. Jeong, A.-L. Barabási. Modeling the Internet's large-scale topology. *Proc. Natl. Acad. Sci.* USA 99 (2002) 13382.
25. S.J. Young and E.R. Scheinerman. Random dot product graph models for social networks. In: Proceedings of WAW 2007, A. Bonato and F.R.K. Chung (Eds.). Springer LNCS 4863:138–149 (2007).
26. R. Xulvi-Brunet and I.M. Sokolov. Evolving networks with disadvantaged long-range connections. *Phys. Rev. E* 66, 026118 (2002).