19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Piecewise-Linear Distance-Dependent Random Graph Models

A.H. Dekker

Defence Science and Technology Organisation, Australia
Email: dekker@acm.org

Abstract: In this paper, we propose a form of random graph (network) model in which the probability of an edge (link) is dependent on a real-valued function on pairs of vertices (nodes). In general, we expect this function to satisfy the triangle inequality, and hence to take the form of a distance metric, although this is not essential. It may reflect physical distances between vertices, dissimilarity measures, or other functions.

We relate the distance metric $d(x,y)$ to the probability $p(x,y)$ of an edge existing between vertices x and y using a piecewise-linear function defined by the triple of parameters (q, a, b):

$$\begin{aligned} p(x,y) &= q, && \text{if } d(x,y) \le a \\ &= 0, && \text{if } d(x,y) > b \\ &= q\,(b - d(x,y)) \,/\, (b - a), && \text{if } a < d(x,y) \le b \end{aligned}$$

A number of important random graph models (Erdős-Rényi graphs, geometric random graphs, and random graphs with a hard distance limit on edges) are special cases of this definition. We use maximum likelihood estimation (MLE) to derive the best parameter triple (q, a, b) for a specific graph. The paper applies this model to a number of example networks, and discusses its practical utility. Figure 1 shows an example network of this type.

A simple SIR disease simulation illustrates the utility of the model for simulation purposes, in exploring a range of network topologies.

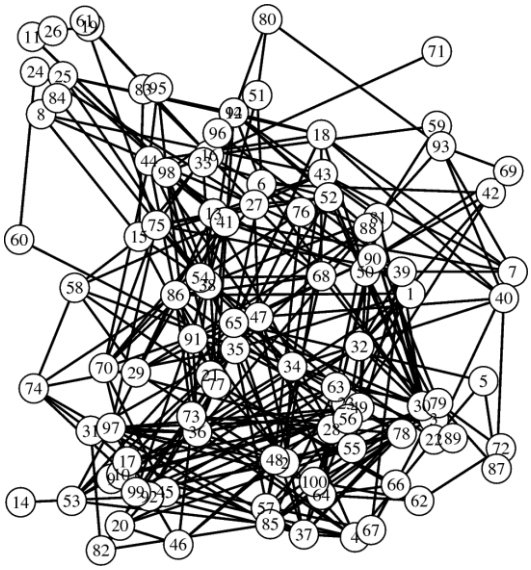

Figure 1. A random graph with a hard limit of 200 on edge lengths. The longest edge length is 199.97. This graph is generated from the parameters (q = 0.164, a = 200, b = 200).

Keywords: *random graph, distance, network, maximum likelihood estimation*

1. INTRODUCTION

In this paper, we consider random graph (network) models where the probability of an edge (link) is dependent on a real-valued function on pairs of vertices (nodes). In general, we expect this function to satisfy the triangle inequality, and hence to take the form of a distance metric, although this is not essential. This real-valued function could reflect *physical distance*, where vertices have an associated location in two-dimensional or three-dimensional space, as in Figure 2. On the other hand, it might indicate distance in a *conceptual space*, where vertices have an *n*-dimensional position resulting from principal components analysis on attributes (Dekker, 2005). The function could also be a *dissimilarity measure* on vertex attributes, or it could reflect a *latent space* (Hoff *et al*., 2002), as in Figure 3, in which vertices are embedded in *n*-dimensional space using a spring-embedding (or, equivalently, multi-dimensional scaling) process (Brandes, 2001). In this latter case, the latent space provides a "summary" of the graph's topological structure.

Our approach is in contrast to *p** models (Carrington *et al*., 2005), in which each pair of distinct vertices corresponds to a binary random variable, and a network of correlations links the random variables. In our approach, all such correlations are incorporated within the distance function. Ours is a less general approach, though adequate for our purpose of generating "plausible" networks for simulation purposes. It also has the advantage of allowing for much simpler analysis.

We relate the distance function $d(x,y)$ to the probability $p(x,y)$ of an edge between vertices x and y using a piecewise-linear function defined by the triple of parameters (q, a, b):

$$\begin{aligned} p(x,y) &= q, && \text{if } d(x,y) \le a \\ &= 0, && \text{if } d(x,y) > b \\ &= q\,(b - d(x,y)) / (b - a), && \text{if } a < d(x,y) \le b \end{aligned}$$

That is, the probability of an edge varies across three distance zones:

- an Erdős-Rényi zone, $d(x,y) \le a$, with a fixed probability q of an edge;
- a forbidden zone, $d(x,y) > b$, where edges do not occur; and
- a transition zone, $a < d(x,y) \le b$, which interpolates linearly between these extremes.

Additional parameters would allow for a nonlinear transition zone, but in real-world networks we have examined, such a step does not appear necessary.

We will write E for the set of edges in the graph, and $\bar{E}$ for the non-edges: pairs of distinct vertices (x,y) which are not edges. We write $P = E \cup \bar{E}$ for the set of all distinct pairs of vertices. We use subscripts and superscripts to express distance restrictions on these sets. For example,

$$E_{\le a} = \{(x,y) \mid (x,y) \in E,\ d(x,y) \le a\}$$

is the set of edges of length at most a. We write $d(S)$ for the multiset of values $d(x,y)$, where $(x,y) \in S$.

Our approach resembles that of Hoff *et al*. (2002), where $p(x,y)$ depends on $d(x,y)$ via a logit function. However, that approach implies that we always have $0 < p(x,y) < 1$, with $p(x,y)$ varying smoothly over the

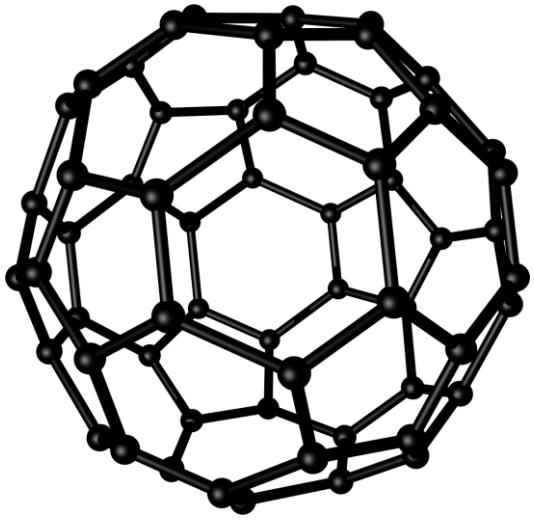

Figure 2. The "soccer-ball" graph (a Cayley graph of the group A_5), embedded in three-dimensional space using the eigenvectors of the first three nonzero eigenvalues.

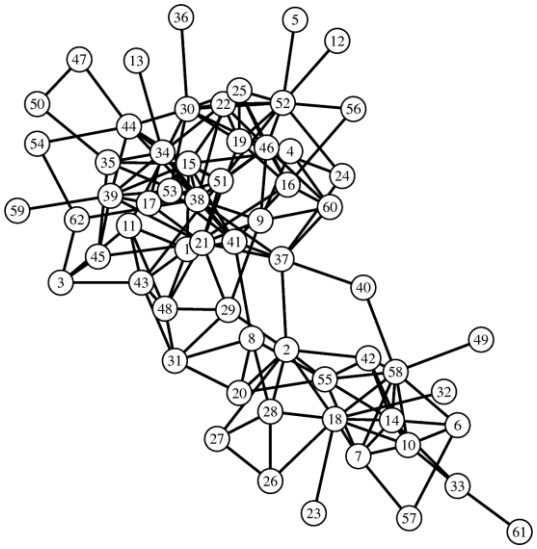

Figure 3. An association network between 62 dolphins living off Doubtful Sound, NZ (Lusseau *et al*., 2003), embedded in 2-dimensional space using spring-embedding/multi-dimensional scaling.

full range of distances. In contrast, we have chosen the piecewise-linear model to allow for several important cases of graphs where this does not hold. For example:

- If $a = b \geq \max(d(P))$ and $q < 1$, we have ordinary Erdős-Rényi random graphs (Bollobás, 2001), where the probability of an edge is q for all pairs of vertices.
- If $\max(d(E)) \leq a = b < \max(d(P))$ and $q < 1$, we have a random graph with a hard limit on edge lengths. We have previously used such graphs as models of sensor networks, where the length restriction is a consequence of physical limitations (Dekker and Skvortsov, 2009).
- If $\max(d(E)) \leq a = b < \max(d(P))$ and $q = 1$, we have a deterministic grid network, in which all pairs of vertices at most a distance a apart are linked with an edge, and no pairs of vertices further apart than this limit are so linked. For example, the "soccer-ball" graph in Figure 2 has two kinds of edges, of length 72 and 82 respectively (scaled to arbitrary units), with the shortest non-edge vertex pair being 116 units apart. This graph can therefore be modelled with $q = 1$ and $a = b = 100$. When the vertices are assigned random positions in space, this becomes the *geometric random graph model*, which has applications to protein interactions (Cannataro *et al*., 2010).

In most cases, however, $q < 1$, $a < b$, and $b > \max(d(E))$. Many real-world graphs can be described by this model, such as the graph in Figure 3. If we can find parameters (q, a, b) describing a particular real-world graph, then realistic alternate topologies can be generated from the associated probability function $p(x,y)$. This is valuable in conducting simulation experiments of networked behaviour, such as those of Dekker (2007).

2. PARAMETER ESTIMATION

Given an arbitrary graph, the parameter estimation problem is to find the parameter triple (q, a, b) which best matches the data. We simplify this problem further by treating q as a derived parameter, with:

$$q = |E_{\leq a}| \,/\, |P_{\leq a}|, \text{ if } |P_{\leq a}| > 0$$
$$= 1, \text{ if } |P_{\leq a}| = 0$$

This ensures that the $P_{\leq a}$ zone has the required Erdős-Rényi behaviour. The parameter estimation problem therefore reduces to finding the best pair (a, b), where $0 \leq a \leq b$. We do this using maximum likelihood estimation (MLE). The likelihood of a pair (a, b) will be nonzero if $a = b \geq \max(d(E))$, or alternatively if $a < b$ and $b > \max(d(E))$. The logarithm of the likelihood can be obtained by the sum of $\log(p(x, y))$ over all edges (x, y) in E, plus the sum of $\log(1 - p(x, y))$ over all non-edges (x, y) in $\bar{E}$. This simplifies to:

$$|E| \log q + |\bar{E}_{\leq a}| \log (1 - q) + \Sigma f(E_{>a,\leq b}) + \Sigma g(\bar{E}_{>a,\leq b}) - |E_{>a,\leq b}| \log (b - a)$$

where

$$f(x, y) = \log (b - d(x,y))$$
$$g(x, y) = \log (1 - q\,(b - d(x,y)) \,/\, (b - a))$$

In general this cannot be maximised analytically, although it is easy to see that if we enforce $a = b$, with $a \geq \max(d(E))$, this simplifies to:

$$|E| \log |E| + |\bar{E}_{\leq a}| \log (1 - |E| \,/\, (|E| + |\bar{E}_{\leq a}|)) - |E| \log (|E| + |\bar{E}_{\leq a}|)$$
$$= |E| \log |E| + |\bar{E}_{\leq a}| \log |\bar{E}_{\leq a}| - (|E| + |\bar{E}_{\leq a}|) \log (|E| + |\bar{E}_{\leq a}|)$$

which is maximised by minimising $|\bar{E}_{\leq a}|$, i.e. by choosing $a = \max(d(E))$.

For the general case, relatively simple search strategies suffice to find the pair (a, b) with the greatest likelihood (or equivalently, the greatest logarithm of the likelihood), even though the log-likelihood function is not convex everywhere. Our implementation uses hill-climbing via a simple genetic algorithm (Goldberg, 1989). This approach gives the same results as brute force search, although more quickly. It also succeeds in recovering, for example, an optimal model for the graph in Figure 2.

3. EXAMPLES

Figure 3 shows an association network between dolphins living off Doubtful Sound, New Zealand (Lusseau *et al*., 2003), embedded in 2-dimensional space using a spring-embedding/multi-dimensional scaling process (Brandes, 2001).

Figure 4 shows the corresponding histogram of vertex-pair distances, binned in steps of 10. Edges (E) are shown in green, and non-edges ($\bar{E}$) in white. Dots show the probability of an edge in each distance bin (measured on the right-hand scale). The solid line shows the probability $p(x, y)$ for the model q = 0.5, a = 62.71, b = 106.42 obtained by MLE estimation.

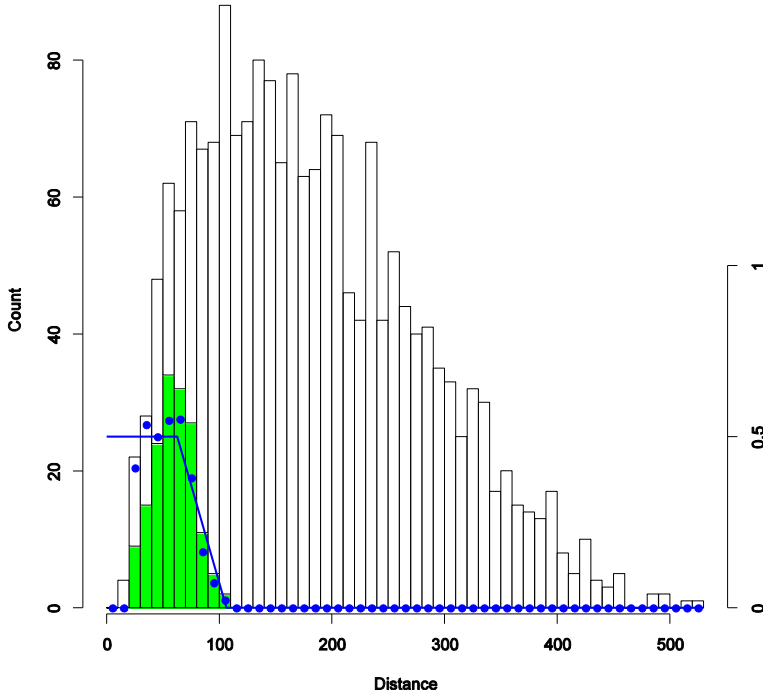

Figure 4. Histogram of edges (green) and non-edges (white) for the dolphin network in Figure 3, binned by distance in steps of 10. Dots show the probability of an edge in each bin (measured on right-hand scale), while the line shows $p(x, y)$ for the MLE-estimated model (0.5, 62.71, 106.42).

We used the MLE-estimated model (0.5, 62.71, 106.42) to generate 240 random graphs with the same number of edges (159) as Figure 3. One example is shown in Figure 5. We are interested in whether global topological properties of the graph are preserved in these randomly generated alternatives. In particular, we are interested in the mean path length, which is also known as the "average distance" – although this refers to path lengths in the graph, not to $d(x, y)$. We are also interested in the clustering coefficient (Watts and Strogatz, 1998), which measures the prevalence of triangles (see Figure 9). For the graph in Figure 3, the mean path length is 3.36 and the clustering coefficient is 0.26. The mean values of these properties for the randomly generated set of 240 graphs were 3.67 (standard deviation 0.26) and 0.20 (s.d. 0.03) respectively, which represents a fairly close match. In other words, graphs like Figure 5, randomly generated from the piecewise-linear distance-dependent random graph model (0.5, 62.71, 106.42), could sensibly replace the graph in Figure 3 for simulation purposes. Although these graphs differ in detail, their global topological properties are similar. In other words, the vertex positions in Figure 3, plus the parameter triple, provide considerable information about the topology of the graph.

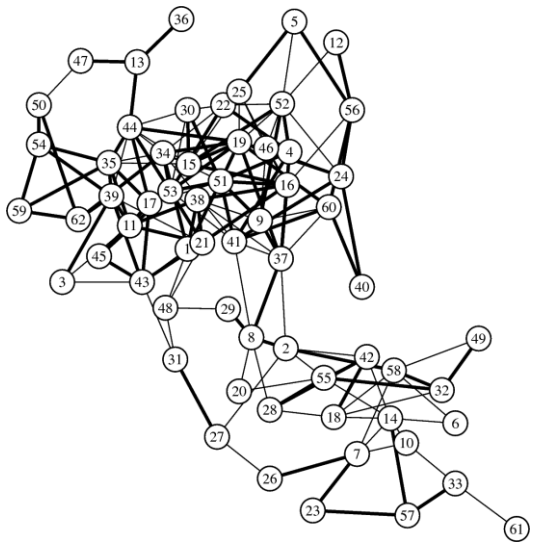

Figure 5. Randomly generated dolphin network, using the parameters (0.5, 62.71, 106.42). Thin lines show edges shared with the original graph (Figure 2), while thick lines show new edges.

Figure 1 shows a random graph with a hard limit on edge lengths. We have previously used such graphs as models of sensor networks, where the length restriction is a consequence of physical limitations (Dekker and Skvortsov, 2009). Figure 6 shows the associated histogram. The model (0.164, 200, 200) used to generate the graph is recovered by the MLE estimation process (even though this process did not assume $a = b$).

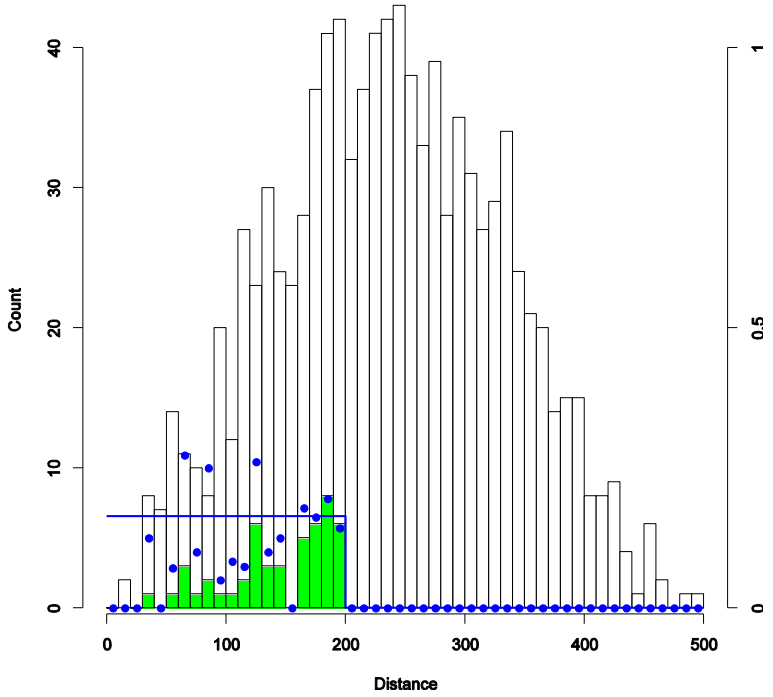

Figure 6. Histogram of edges (green) and non-edges (white) for the graph in Figure 1, binned by distance. Dots show the probability of an edge in each bin (on right-hand scale). The line shows the model (0.16, 200, 200) used to generate the graph, which is also recovered by MLE.

4. DEGREE-BASED MODELS

Figure 7 shows a graph of airline routes in the US, with vertex location based on the latitude and longitude of airports. This is a "scale-free" (preferential-attachment) graph (Albert and Barabási, 2002; Barabási, 2002), with vertex degrees following roughly a power-law distribution. As a consequence the maximum degree (130) is substantially greater than the mean degree (11.04). To model graphs of this kind, we allow edge probabilities to also depend on the degree δ_x of vertices x:

$$\begin{aligned} p(x,y) &= q\,\delta_x\,\delta_y, && \text{if } d(x,y) \le a \\ &= 0, && \text{if } d(x,y) > b \\ &= q\,\delta_x\,\delta_y\,(b - d(x,y)) / (b - a), && \text{if } a < d(x,y) \le b \end{aligned}$$

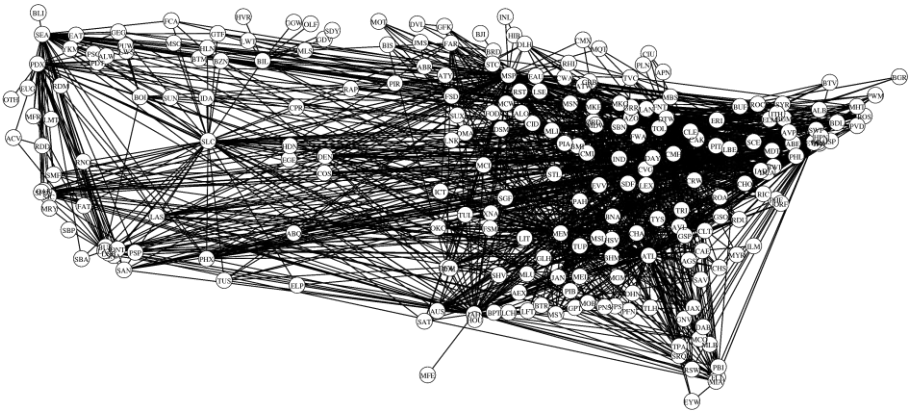

Figure 7. A graph of airline routes in the US, with 235 vertices and 1297 edges. The mean path length is 2.32 and the clustering coefficient is 0.56. The degree distribution roughly follows a power law.

For the graph in Figure 7, the MLE model (0.07, 27.66, 1078.50) produces graphs with a similar mean path length – over a sample of 450 randomly generated graphs, a mean of 2.48 (s.d. 0.02). However, the clustering coefficient produced is much lower – a mean of 0.26 (s.d. 0.02). This is because edges in Figure 7 are not highly localised. Preservation of clustering coefficients relies on the following causal chain:

- (x, y) and (y, z) are edges;
- therefore $d(x, y)$ and $d(y, z)$ are likely to be low (assuming edges are localised);
- therefore $d(x, z)$ is likely to be low (assuming the triangle inequality or something similar);
- therefore (x, z) is likely to be an edge, forming a triangle with (x, y) and (y, z).

This kind of reasoning does not hold for the graph in Figure 7, or for other graphs where edges are not localised, as they were in Figure 1 and Figure 3.

5. BIPARTITE GRAPHS

Figure 8 shows the well-known "Davis' Southern Club Women" bipartite graph (Davis *et al*. 1941; Breiger, 1974). For bipartite graphs, analysis proceeds exactly as above, but we assume the graph is bicoloured, and require both edges E and non-edges $\bar{E}$ to involve vertices of different colours.

Clustering coefficients, however, are not appropriate for bipartite graphs, since triangles cannot occur. We therefore suggest an analogous measure, the *diamond coefficient*, defined as the mean of C_x over all vertices x, where C_x is the probability that, given edges (x, y) and (x, z), there also exist edges (y, u) and (z, u) for some vertex u. Figure 9 illustrates this definition, together with that of the traditional clustering coefficient.

An alternate measure to the diamond coefficient is that suggested by Latapy *et al*. (2008), who use the probability that, given three edges, a fourth edge exists which completes the diamond.

The graph in Figure 8 has a mean path length of 2.29 and a diamond coefficient of 0.96. The model produced by MLE was (0.94, 135.88, 168.79), and a set of 500 graphs generated from this model had similar mean path lengths (mean 2.31, s.d. 0.03). All generated graphs had diamond coefficients identical to the original. This reflects the strong degree of edge localisation in Figure 8.

6. SIMULATION EXAMPLE

As an example of the use of piecewise-linear distance-dependent graph models for simulation, we constructed a range of (q, a, b) graphs for a given set of 100 nodes. In each case, q was chosen so as to give an average degree of 4. The graphs were then

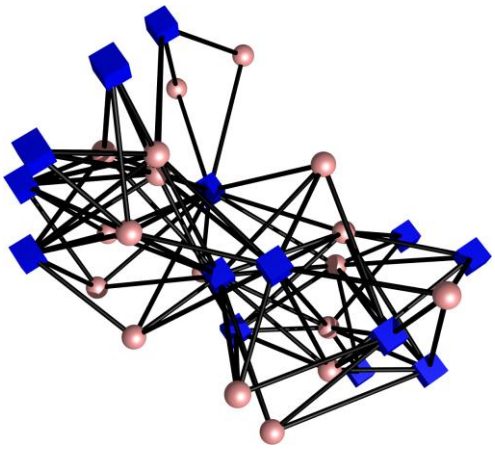

Figure 8. The bipartite graph of Davis' 18 Southern Club Women (pink spheres) and the 14 events they helped organise (blue cubes), embedded in 3-dimensional space using spring-embedding/multi-dimensional scaling (Davis *et al*. 1941; Breiger, 1974).

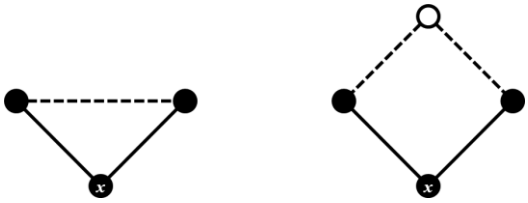

Figure 9. The clustering coefficient (left) is the probability (averaged over all vertices x) that the partial triangle formed by two adjacent edges is completed by a third edge (dashed). By analogy the diamond coefficient (right) for bipartite graphs is the probability (averaged over all vertices x) that the partial diamond formed by two adjacent edges is completed by two additional edges (dashed).

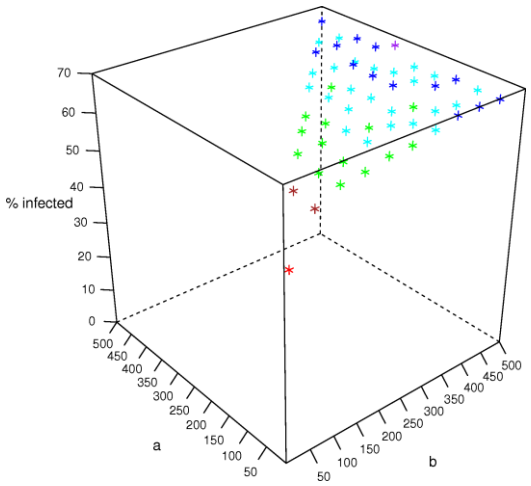

Figure 10. Results for SIR simulation on 100-node (q, a, b) graphs as a function of a and b. The spread of infection ranges from 44% at $a = b = 50$ to 68% at $a = 300$, $b = 500$. The variation is primarily due to the value of b in this case.

used as input to an SIR disease model (Giesecke, 2002; Dekker, 2008) with one initially infected node and a 0.5 probability of a node infecting a given neighbour. Figure 10 shows the spread of infection (averaged over 10,000 runs) for several $a \leq b$ pairs. The variation is primarily due to b, since high values of b allow for a small number of long-distance edges, which facilitate the spread of infection (Watts and Strogatz, 1998). A simple cubic ($40 + 0.20\ b - 5.6\times10^{-4}\ b^2 + 5.2\times10^{-7}\ b^3$) predicts 83% of the variance. Visualising behaviour as a function of a and b, as in Figure 9, is a simple but effective way of showing the impact of edge length.

7. DISCUSSION

In this paper, we have outlined a model for piecewise-linear distance-dependent random graphs where, based on the distance $d(x,y)$ between vertices x and y, there is:

- an Erdős-Rényi zone, $d(x,y) \leq a$, with a fixed probability q of an edge;
- a forbidden zone, $d(x,y) > b$, where edges do not occur; and
- a transition zone, $a < d(x,y) \leq b$, which interpolates linearly between these extremes.

This generalises Erdős-Rényi random graphs, geometric random graphs, and random graphs with a hard distance limit on edges. The model extends easily to bipartite graphs.

When model parameters are derived from a graph with spatially localised edges, a class of random graphs is produced where the overall topology – as measured by the mean path length and clustering or diamond coefficient – resembles the original graph. This can provide a pool of different but similar graphs for use in simulations of networked behaviour.

A simple SIR simulation illustrates the utility of the model for exploring a range of network topologies.

REFERENCES

Albert, R. and Barabási, A.-L. (2002), "Statistical mechanics of complex networks," *Reviews of Modern Physics*, **74**, 47–97.

Barabási, A.-L. (2002), *Linked: The New Science of Networks*, Perseus Publishing, Cambridge, MA.

Bollobás, B. (2001), *Random Graphs*, 2nd edition Cambridge University Press, Cambridge, UK.

Brandes, U. (2001), "Drawing on Physical Analogies," in M. Kaufmann and D. Wagner, eds., *Drawing Graphs: Methods and Models*, 71–86, Springer Verlag (LNCS 2025).

Breiger, R. (1974), "The Duality of Persons and Groups," *Social Forces*, **53**(2): 181–190.

Cannataro , M., Guzzi, P.H., and Veltri, P. (2010), "Protein-to-protein interactions: Technologies, databases, and algorithms," *ACM Computing Surveys*, **43**(1): Article 1.

Carrington, P., Scott, J., and Wasserman, S. (2005), *Models and Methods in Social Network Analysis*, Cambridge University Press.

Davis, A., Gardner, B., and Gardner, M. (1941), *Deep South: A Social Anthropological Study of Caste and Class*, University of Chicago Press (re-issued 2009 by University of South Carolina Press).

Dekker, A.H. (2005), "Conceptual Distance in Social Network Analysis," *Journal of Social Structure*, **6**(3): www.cmu.edu/joss/content/articles/volume6/dekker/

Dekker, A.H. (2007), "Studying Organisational Topology with Simple Computational Models," *Journal of Artificial Societies and Social Simulation*, **10**(4): jasss.soc.surrey.ac.uk/10/4/6.html

Dekker, A.H. (2008), "Network Effects in Epidemiology," *Proc. SimTecT 2008*, pp 39–44.

Dekker, A.H. and Skvortsov, A.T. (2009), "Topological Issues in Sensor Networks, " in Anderssen, R.S., Braddock, R.D., and Newham, L.T.H., eds., *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, Modelling and Simulation Soc. of Aust. and NZ and Int'l Assoc. for Mathematics and Computers in Simulation, 952–958.

Giesecke, J. (2002), *Modern Infectious Disease Epidemiology*, 2nd edition, Arnold.

Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Hoff, P.D., Raftery, A.E., and Handcock, M.S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, **97**(460): 1090–1098.

Latapy, M., Magnien, C., and Del Vecchio, N. (2008) "Basic Notions for the Analysis of Large Affiliation Networks / Bipartite Graphs," arXiv:cond-mat/0611631v1.

Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., and Dawson, S.M. (2003), "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, **54**, 396–405.

Watts, D.J. and Strogatz, S.H. (1998), "Collective dynamics of 'small world' networks," *Nature*, **393**, 440–442.