Random Dot Product Graph Models for Social Networks

Stephen J. Young[1] and Edward R. Scheinerman[2]

[1] Georgia Institute of Technology, Atlanta GA 30332-0160
[2] Johns Hopkins University, Baltimore MD 21218-2682

Abstract. Inspired by the recent interest in combining geometry with random graph models, we explore in this paper two generalizations of the random dot product graph model proposed by Kraetzl, Nickel and Scheinerman, and Tucker [1,2]. In particular we consider the properties of clustering, diameter and degree distribution with respect to these models. Additionally we explore the conductance of these models and show that in a geometric sense, the conductance is constant.

1 Introduction

With the ubiquity and importance of the Internet and genetic information in medicine and biology, the study of complex networks relating to the Internet and genetics continues to be an important and vital area of study. This is especially true for networks such as the physical layer of the Internet, the link structure of the world wide web, and protein-protein and protein-gene interaction networks. Due to the size of these networks [3] and the difficulty of determining complete link information [4,5] a significant amount of research has gone into finding models that match observed properties of these graphs in order to empirically (via simulation) and theoretically understand and predict properties of these complex networks. There are three models that, together with their variations, are the core models for these complex networks [6]. The configurational model and its variants attempt to generate complex networks by specifying the degree sequence and creating edges randomly with respect to that degree sequence. On the other hand, the Barabási-Albert preferential attachment model attempts to model the process by which the network grows, specifically, it posits that vertices with high degree are more likely to increase in degree when a new vertex is added to the network. In a similar vein, the copying model [7,8], also attempts to model the growth process of a complex networks. However, the copying model takes the more distinctly biological viewpoint of replication of existing nodes combined with mutation. All three of these types of models have had success in reproducing the hallmark features of complex networks, namely a power-law degree distribution, a diameter that grows slowly or is constant with the size of the graph, and one of several clustering properties; see [6,9] for a collection of such results.

However, there are many other aspects of complex networks that fail to be captured by these models, for example non-uniform assortativity [10] and the

A. Bonato and F.R.K. Chung (Eds.): WAW 2007, LNCS 4863, pp. 138–149, 2007.

existence of directed cycles, among others. Thus there is considerable interest in new models for complex networks that exhibit a power-law like degree sequence, small diameter, and clustering, and are different enough from the three main model classes to exhibit other properties of complex networks that are not exhibited by the current models. One potential method to create new models is to incorporate geometry into already existing models. Flaxman, et al. used geometry coupled with the preferential attachment model to create a model that generates a random power-law graph that has small separators [11].

Taking this idea one step further, one can add semantic information to an already existing model. One such model is the random dot product graph model applied by Caldarelli, et al. and Azar, et al. [12,13] and formalized by Kraetzl, Nickel, Scheinerman, and Tucker [1,2]. In their work they assign to each vertex a vector in $\mathbb{R}^d$ and then any edge is present with probability equal to the dot product of the endpoints. Thus, thinking of the vertices as members of a social network, the vectors together with the dot product encode semantically the idea of differing "interests" and varying levels of "talkativeness." We discuss the two natural generalizations of the random dot product graph model proposed by Kraetzl, et al, specifically, we remove the restrictions on the vectors imposed in their earlier work and develop directed generalization. First we briefly outline in Sect. 2 their model and the known results on diameter, clustering and degree distribution in order to provide a framework for the rest of this paper. We then present the two natural generalizations of the random dot product model. In Sect. 3 and Sect. 4 we demonstrate that an arbitrarily large fraction of the graph has constant diameter and that both the undirected and directed models demonstrate clustering. We derive in Sect. 5 explicit formulas for the degree sequence leading to a super-linear number of edges, which is consistent with recent results of Leskovec, Kleinberg and Faloutsos [14]. Finally, in Sect. 6 we turn our attention to conductance. We show that any small separators present are essentially non-semantic and leave open the question of general conductance. In Sect. 7 we discuss some areas for future work.

2 Model Specification

Kraetzl, Nickel and Scheinerman develop a new family of random graph model for social networks based on the dot product. In particular, they consider in detail the following model. Each vertex v is independently assigned a random vector, W_v, in $\mathbb{R}^d$, where each coordinate is independently and identically distributed as $\frac{1}{\sqrt{d}}\mathcal{U}^\alpha[0,1]$; that is a scaled copy the uniform distribution on $[0,1]$ to the α power. Then each edge $\{u,v\}$ is present independently with probability $\langle W_v, W_u\rangle$. They go on to show that the resulting graph G, has the following properties for $d=1$:

1. The giant component of G has diameter almost surely at most 6 as $n \to \infty$.
2. For all vertices u, v, and w,

$$\mathbb{P}\left(u \sim w \mid u \sim v \sim w\right) = \left(\frac{\alpha+1}{2\alpha+1}\right)^2 > \mathbb{P}\left(u \sim w\right) = \frac{1}{(\alpha+1)^2} \ . \qquad (1)$$

3. The expected number of vertices of degree k on a n vertex graph generated in this manner is
$$\frac{1}{k!\alpha}(1+\alpha)^{\frac{1}{\alpha}}\Gamma\left(\frac{1}{\alpha}+k\right)n^{\frac{\alpha-1}{\alpha}} \quad . \tag{2}$$

They proceed to show that for higher dimensions the probability of an arbitrary edge is independent of the dimension, but the degree distribution develops a "bend" in the power law. That is, the slope of the log-log plot of the degree distribution in numerical studies (and confirmed analytically for $d = 2$) decreases sharply for some given degree, which they conjecture to be $n/(d\alpha + d)$.

We consider the two natural generalizations of this model, one undirected and one directed, and show that they behave similarly to the model described by Kraetzl, Nickel and Scheinerman and resolve some of the higher dimensional questions posed regarding the nature of clustering and the diameter in the model.

First we consider the undirected generalization. Let $\mathbf{W}$ be a random variable on a $\mathbb{R}^d$ such that if W_i and W_j are distributed as $\mathbf{W}$, $\mathbb{P}(\langle W_i, W_j\rangle \in (0,1)) = 1$. Then we define $G(\mathbf{W}, n)$ as the graph on n vertices where each vertex v is assigned a vector W_v distributed as $\mathbf{W}$ and each edge $\{u, v\}$ is present independently with probability $\langle W_u, W_v\rangle$. It is clear from this construction that the restriction on the nature of the distribution $\mathbf{W}$ is necessary in order to guarantee that the inner products are all valid, nontrivial probabilities. When a distribution satisfies this condition, we shall say that it satisfies the inner product condition. Note that the inner product condition implies that $\mathbb{P}(\|\mathbf{W}\| < 1) = 1$, and guarantees that there is always some probability of an edge appearing (or not appearing) between any two pairs of vertices. Although it may seem more natural to allow for 0 or 1 inner products, precluding these values simplifies the analysis by forbidding pathological and uninteresting cases that can come about when there is a positive probability of guaranteeing or forbidding an edge.

The natural generalization of $G(\mathbf{W}, n)$ is to consider a directed graph with similar properties. Suppose $(\mathbf{X}, \mathbf{Y})$ is a pair of distributions on $\mathbb{R}^d \times \mathbb{R}^d$ such that if X_u is distributed as $\mathbf{X}$ and Y_v is distributed as $\mathbf{Y}$, $\mathbb{P}(\langle X_u, Y_v\rangle \in (0,1)) = 1$. We will abuse terminology slightly and say that such a $(\mathbf{X}, \mathbf{Y})$ pair satisfies the inner product condition. Then we consider the random directed graph $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ as the graph on n vertices, where each vertex v is assigned a pair of vectors (X_v, Y_v) and each directed edge (u, v) is present independently with probability $\langle X_u, Y_v\rangle$. Again, the inner product condition is a natural condition driven by the necessity for the quantity associated to an arc being a probability. Note that it is clearly not necessary for either of $\mathbf{X}$ or $\mathbf{Y}$ to have bounded norm, however we believe that the nature of those distributions such that $(\mathbf{X}, \mathbf{Y})$ satisfies the inner product condition and has unbounded norm are so pathological as to be uninteresting. Thus, for the remainder of this paper we assume that there is some compact set K such that $\mathbb{P}(\mathbf{X} \in K) = \mathbb{P}(\mathbf{Y} \in K) = 1$. Note as well that for clarity of presentation, we will abuse notation and say that a vertex belongs to a region R whenever its assigned vector(s) lie in that region.

We observe that $G(\mathbf{W}, n)$ generalizes both the Erdős-Rényi model and a version of the configurational random graph model. The first is achieved by letting

$\mathbf{W}$ be a constant random variable. Then it is clear that the model under consideration is just the Erdős-Rényi model with parameter $\langle \mathbf{W}, \mathbf{W} \rangle$. Also note that this holds for $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ by letting both $\mathbf{X}$ and $\mathbf{Y}$ be constant. Now by letting $d = 1$ and $\mathbb{P}(\mathbf{W} = k/c)$ be proportional to $k^{-\alpha}$, where c is a normalizing constant, we have a model that generalizes a randomized configurational model.

In addition to generalizing the Erdős-Rényi and configurational models, there is a natural interpretation of the vectors and the interaction of those vectors in the (directed) random dot product graph model. By considering each component of the vector associated with a vertex as a property or interest of that vertex, we may interpret the value of the component in a natural way. Furthermore, recent research into the nature of links in the blogosphere, specifically the Live Journal networks, have shown that a significant percentage of links can be explained by properties of the blog, such as the location of the author, interest lists, age, gender, etc. [15]. This interpretation of random dot product graphs provides a ready-made collection of tools for creating distributions by applying previous research into the singular value decomposition and related methods for feature extraction.

Just as representing entities as vectors, or pairs of vectors, is a natural idea, we feel that the inner product is a natural way of encapsulating two primary barriers to "linking". More explicitly, two websites are unlikely to have a direct link if their topics are completely unrelated, this corresponds to their vectors having a large angle between them in the dot product graph representation. On the other hand, if two websites have nearly identical topics, they still may not be linked due to the selectivity of one of the websites. That is, if one of the websites doesn't link to many things overall, then no matter how close another website's interests are there is still a significant barrier to "linking". The inner product encapsulates both these barriers in that both the angle between the vector and the norm of the vectors impact the inner (dot) product.

3 Diameter of "Giant" Component

In this section, we show that an arbitrarily large fraction of the graph generated by $G(\mathbf{W}, n)$ almost surely forms a connected graph with diameter at most 5. In a slight abuse of standard terminology, we will refer to this arbitrarily large fraction of the graph as the "giant" component. A key step in the proof of the diameter of the "giant" component for $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ is the following lemma, which generalizes the result on the diameter of the Erdős-Rényi random graph model.

Lemma 1. *Let D be a directed random graph on v vertices such that each directed edge is present independently with probability at least p. Then D is almost surely strongly connected with directed diameter 2.*

Proof. Consider some pair of vertices, u and v. The probability that there is not a directed path of length at most 2 from u to v is at most $(1-p^2)^{|V(D)|-2}(1-p)$. Thus the probability that u and v are not strongly connected by paths of length at most 2 is at most $1-(1-(1-p^2)^{n-2}(1-p))^2$. But then, the expected number

of such pairs that are not strongly connected by paths of length at most 2 is at most

$$n(n-1)(2(1-p^2)^{n-2}(1-p) - (1-p^2)^{2n-4}(1-p)^2) \tag{3}$$

which approaches 0 as $n \to \infty$. Thus D is almost surely strongly connected with directed diameter at most 2 [16].

We will denote by $B(c;r)$ (respectively $\overline{B}(c;r)$) the open (respectively closed) ball of radius r centered at c.

Theorem 1. *Let $\mathbf{W}, \mathbf{X}, \mathbf{Y}$ be distributions on $\mathbb{R}^d$ such that $\mathbf{W}$ and $(\mathbf{X}, \mathbf{Y})$ satisfy the inner product condition. Further assume that there is some compact region K such that $\mathbf{X}$ and $\mathbf{Y}$ lie inside K almost surely. Then an arbitrarily large fraction of $G(\mathbf{W}, n)$ is connected with diameter 5 and an arbitrarily large fraction of $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ is strongly connected with directed diameter at most 5.*

We prove only the undirected case here as the directed case follows a similar but more complicated argument.

Proof. We may assume without loss of generality that $\mathbf{W} \in \overline{B}(0;1)$. Letting $0 < \delta < \frac{1}{4}$, choose $\epsilon > 0$ such that $\mathbb{P}(\mathbf{W} \in B(0;\epsilon)) < \delta$. Then let A be the closed annulus $\overline{B}(0;1) - B(0;\epsilon)$. For all $\alpha \in A$, choose

$$r_\alpha \in \left\{ r > 0 | \, \forall x, y \in B(\alpha; r), x^T y > \frac{\epsilon^2}{4} \right\} \ , \tag{4}$$

which is non-empty by the continuity of the inner product. Then $\cup_{\alpha \in A} B(\alpha; r_\alpha)$ is an open cover of the compact set A with some finite subcover, say $\{B(\alpha_i; r_{\alpha_i})\}$.

Fix i such that $\mathbb{P}(\mathbf{W} \in B(\alpha_i; r_{\alpha_i})) \neq 0$. Then, as $n \to \infty$, there are almost surely infinitely many vertices that lie in $B(\alpha_i; r_{\alpha_i})$. It then follows from a result of Erdős and Renyi, since the probability of every edge is at least $\frac{\epsilon^2}{4}$ and for fixed $\{W_v\}$ each edge is present independently, the graph induced by $B(\alpha_i; r_{\alpha_i})$ has diameter at most 2, almost surely. Clearly, if $\mathbb{P}(\mathbf{W} \in B(\alpha_i; r_{\alpha_i})) = 0$, then there are almost surely no vertices in that region, and moreover those regions do not affect the diameter of $G(\mathbf{W}, n)$.

Now consider two regions $\mathcal{R}_i = B(\alpha_i; r_{\alpha_i})$ and $\mathcal{R}_j = B(\alpha_j; r_{\alpha_j})$ occurring with positive probability. There is a naturally defined probability measure on $\mathcal{R}_i \times \mathcal{R}_j$. Furthermore, since $\mathbb{P}(W_i^T W_j = 0) = 0$, there exist $\hat{\epsilon}, \hat{\delta} > 0$ such that $\mathbb{P}\left(W_i^T W_j > \hat{\delta} \mid W_i \in \mathcal{R}_i, W_j \in \mathcal{R}_j\right) > \hat{\epsilon}$. But, since $\hat{\delta}$ and $\hat{\epsilon}$ are independent of n, and since $\mathcal{R}_i$ and $\mathcal{R}_j$ almost surely contain an infinite number of vertices; there is almost surely an edge between the regions. Now given vertices $u \in \mathcal{R}_i$ and $v \in \mathcal{R}_j$, there is almost surely an edge e between $\mathcal{R}_i$ and $\mathcal{R}_j$, a path of length 2 from u to e, and a path of length 2 from e to v. Thus, for any pair of vertices in A there is almost surely a path of length at most 5 between them. But A asymptotically contains $(1-\delta)n$ vertices, and since δ was arbitrary, A contains an arbitrarily large fraction of the vertices.

4 Clustering

In this section, we examine the clustering of $G(\mathbf{W}, n)$ and $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ and find that except in the case of constant random variables, the presence clustering is independent of the random variables. In order to show the clustering results we need the following convexity result, which will allow the use of Jensen's Inequality in the proof of Theorem 2.

Lemma 2. *Let $a, b \in \mathbb{R}^d$. Let $D \subseteq \mathbb{R}^d$ be a region such that for all $x \in D$, $\langle a, x\rangle \in (0,1)$ and $\langle b, x\rangle \in (0,1)$. Then $u\colon D \longrightarrow \mathbb{R}$ defined by $x \longmapsto \langle a, x\rangle \langle b, x\rangle$ is a convex function of x.*

Proof. Let $F : (0,1) \times (0,1) \longrightarrow \mathbb{R}$ be defined by $(x, y) \longmapsto xy$. We note that $\nabla^2 F = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. This matrix, although not positive semi-definite, is positive semidefinite over $[0,1] \times [0,1]$, and hence $F(x, y)$ is convex over its domain [17]. Now note that since $\langle a, x\rangle$ is a real inner product, for any $\lambda \in [0,1]$ and $x, y \in D$, $\langle a, \lambda x + (1-\lambda)y\rangle = \lambda \langle a, x\rangle + (1-\lambda)\langle a, y\rangle$. Thus $\langle a, x\rangle$ is a convex function in x and similarly for $\langle b, x\rangle$. Thus $u(x) = F(\langle a, x\rangle, \langle b, x\rangle)$ is the composition of convex functions and hence is convex.

Theorem 2. *Let $W_v, W_w, W_u, X_u, X_v, X_w, Y_u, Y_v, Y_w$ be independent random variables distributed over $\mathbb{R}^d$, not necessarily identically distributed, such that $\langle W_i, W_j\rangle$ and $\langle X_i, Y_j\rangle$ satisfy the inner product condition for all $i \neq j$. For the undirected graph where each edge $\{i, j\}$ is present with probability $\langle W_i, W_j\rangle$, we have that*

$$\mathbb{P}(u \sim v \mid u \sim v, v \sim w) \geq \mathbb{P}(u \sim v). \tag{5}$$

Now consider the random directed graph where each arc $i \to j$ is present, independently, with probability $\langle X_i, Y_j\rangle$. Then we have that

1. $\mathbb{P}(u \to w \mid u \to v, v \to w) \geq \mathbb{P}(u \to w)$,
2. $\mathbb{P}(u \to w \mid u \to v, w \to v) \geq \mathbb{P}(u \to w)$,
3. $\mathbb{P}(u \to w \mid v \to u, v \to w) \geq \mathbb{P}(u \to w)$, *and*
4. $\mathbb{P}(u \to w \mid w \to v, v \to u) = \mathbb{P}(u \to w)$.

As an immediate corollary, we get that for any set of vertices u, v and w in $G(\mathbf{W}, n)$, we get $\mathbb{P}(u \sim v \mid u \sim v, v \sim w) \geq \mathbb{P}(u \sim v)$ and for any set of vertices u, v and w in $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$

1. $\mathbb{P}(u \to w \mid u \to v, v \to w) \geq \mathbb{P}(u \to w)$,
2. $\mathbb{P}(u \to w \mid u \to v, w \to v) \geq \mathbb{P}(u \to w)$,
3. $\mathbb{P}(u \to w \mid v \to u, v \to w) \geq \mathbb{P}(u \to w)$, and
4. $\mathbb{P}(u \to w \mid w \to v, v \to u) = \mathbb{P}(u \to w)$.

Note that in $G(\mathbf{W}, n)$, equality holds bounds on clustering if and only if $\mathbf{W}$ is a constant random variable.

5 Degree Distribution

We derive, in this section, a set of natural formulas for the degree distributions of both $G(\mathbf{W}, n)$ and $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$. In Sect. 5.1, we discuss the application of these formulas to the construction of specific random models meeting desired degree sequence considerations.

Proposition 1. *Let $G = G(\mathbf{W}, n)$ where $\mathbf{W}$ satisfies the inner product condition and let $D = \overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ where $\mathbf{X}$ and $\mathbf{Y}$ are distributions over $\mathbb{R}^d$ where $(\mathbf{X}, \mathbf{Y})$ satisfies the inner product condition. Then, for a vertex $w \in V(G)$*

$$\mathbb{P}\left(\deg(w) = k\right) = \int \binom{n-1}{k} \left\langle \mathbb{E}\left[\mathbf{W}\right], W\right\rangle^k \left(1 - \left\langle \mathbb{E}\left[\mathbf{W}\right], W\right\rangle\right)^{n-k-1} d\mathbf{W}. \quad (6)$$

Furthermore, for a vertex $v \in V(D)$

$$\mathbb{P}\left(\deg^+(v) = k\right) = \int \binom{n-1}{k} \left\langle \mathbb{E}\left[\mathbf{X}\right], Y\right\rangle^k \left(1 - \left\langle \mathbb{E}\left[\mathbf{X}\right], Y\right\rangle\right)^{n-1-k} d\mathbf{Y} \quad (7)$$

$$\mathbb{P}\left(\deg^-(v) = k\right) = \int \binom{n-1}{k} \left\langle \mathbb{E}\left[\mathbf{Y}\right], X\right\rangle^k \left(1 - \left\langle \mathbb{E}\left[\mathbf{Y}\right], X\right\rangle\right)^{n-1-k} d\mathbf{X}\ . \quad (8)$$

This leads to an immediate result on the density of edges in $G(\mathbf{W}, n)$ and $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$.

Corollary 1. *Let $G = G(\mathbf{W}, n)$ where $\mathbf{W}$ satisfies inner product condition and let $D = \overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$ where $\mathbf{X}$ and $\mathbf{Y}$ are distributions over $\mathbb{R}^d$ where $(\mathbf{X}, \mathbf{Y})$ satisfies the inner product condition. Then $\mathbb{E}\left[|E(G)|\right] = \binom{n}{2} \left\langle \mathbb{E}\left[\mathbf{W}\right], \mathbb{E}\left[\mathbf{W}\right]\right\rangle$ and $\mathbb{E}\left[|E(D)|\right] = n(n-1) \left\langle \mathbb{E}\left[\mathbf{X}\right], \mathbb{E}\left[\mathbf{Y}\right]\right\rangle$.*

This implies that the edge density is $\Omega\left(n^2\right)$, contrary to conventional wisdom regarding complex networks. However we feel that this trade off is acceptable in practice for several reasons. The first being that $\left\langle \mathbb{E}\left[\mathbf{W}\right], \mathbb{E}\left[\mathbf{W}\right]\right\rangle$ and $\left\langle \mathbb{E}\left[\mathbf{X}\right], \mathbb{E}\left[\mathbf{Y}\right]\right\rangle$ are typically small. Furthermore, although the results regarding the diameter of the graph would not hold, one could consider $\mathbf{X}$ and $\mathbf{Y}$ as functions of n and introduce sparsity in that manner. We also note that, particularly for the world wide web, gene-protein networks, and the Internet, it is widely accepted that empirical studies are not capturing all the edges present. Combine this fact with recent work showing that the incompleteness can severely skew some statistics of the data [4,5], and it is plausible that one or more of these networks is not truly sparse. In addition, the recent work of Leskovec, Kleinberg and Faloutsos [14] has shown that for many social networks the number of edges is becoming super-linear in the number of vertices as these networks evolve.

5.1 Sample Distributions

Although it is obvious that not every distribution $\mathbf{W}$ or pair of distributions $(\mathbf{X}, \mathbf{Y})$ can lead to a power law, it is useful to discuss a few means of generating

power law degree distributions. We will focus on the directed model, as Kraetzl, Nickel and Scheinerman have already shown one manner in which to achieve a power law degree distribution for the undirected model [1].

It is natural to consider directed versions of complex network where the in-degrees are distributed as a power law, while the out-degrees tend to be more concentrated, in order to capture situations where the for physical reasons the the out-degree is limited. Thus since we know that the Erdős-Rényi graph model tends to produce a concentrated degree sequence and further if each component is independently distributed as $\frac{1}{\sqrt{d}}\mathcal{U}^{\alpha}(0,1)$ in the undirected random dot product graph model tends to produce a power law, it is natural to attempt to emulate these two in the directed model. Thus, taking each component of $\mathbf{X}$ to be independently distributed as $\frac{1}{\sqrt{d}}\mathcal{U}(0,1)$ (that is, having low variance, similarly to the Erős-Rényi model) and each component of $\mathbf{X}$ to be $\frac{1}{\sqrt{d}}\mathcal{U}^{\alpha}(0,1)$ (and thus having high variance), with $\alpha = 16$, $d = 5$, $n = 10000$, and with 200 trials, yields the average degree distributions shown in Fig. 5.1. Note that this roughly models the desired behavior, in that the out-degree is strongly clustered around a single value and the in-degrees are distributed as a power-law. Further empirical refinement can lead to a closer approximation, and thus through repeated simulation and tuning it is reasonable to assume that this degree distribution and others like it, can be well approximated.

We now note that for any orthonormal matrix Q and any non-zero constant c, $\overrightarrow{G}(\mathbf{X},\mathbf{Y},n) = \overrightarrow{G}(cQ\mathbf{X},\frac{1}{c}Q\mathbf{Y},n)$. Thus we may assume that

1. $\langle \mathbb{E}[\mathbf{X}], e_1\rangle = \langle \mathbb{E}[\mathbf{Y}], e_1\rangle = \sqrt{\langle \mathbb{E}[\mathbf{X}], \mathbb{E}[\mathbf{Y}]\rangle}$,
2. $\langle \mathbb{E}[\mathbf{X}], e_2\rangle \geq 0$,
3. $\langle \mathbb{E}[\mathbf{Y}], e_2\rangle = 0$, and
4. $\langle \mathbb{E}[\mathbf{X}], e_i\rangle = \langle \mathbb{E}[\mathbf{Y}], e_i\rangle = 0$ for $i > 2$.

In particular, combining these observations with (7) and (8), we obtain moments for some components of $\mathbf{X}$ and $\mathbf{Y}$ if $\overrightarrow{G}(\mathbf{X},\mathbf{Y},n)$ satisfies a given degree distribution. However, these moments do not fully characterize $\mathbf{X}$ and $\mathbf{Y}$, but rather limit the space of feasible distributions.

Perhaps more useful from a modeling point of view is the possibility of using Kernel Density Estimators of Hörmann and Leydold [18] to develop estimated distributions for $(\mathbf{X},\mathbf{Y})$. In particular, given a graph G and a vector $\{V_i\}$ for each vertex so that G is "generated" by $\{V_i\}$ under the random dot product graph model, Kernel Density Estimation provides a means to rapidly generate approximate samples from the sample distribution. (For more details on extracting vectors from a given graph see [19,20,21].) Thus using $\{V_i\}$ it is theoretically possible to generate a random graph that "looks-like" G.

6 On the Nature of Bad Cuts

In light of the work of Flaxman, Frieze and Vera which showed that the geometric preferential attachment graph has bad cuts that are due to the geometry

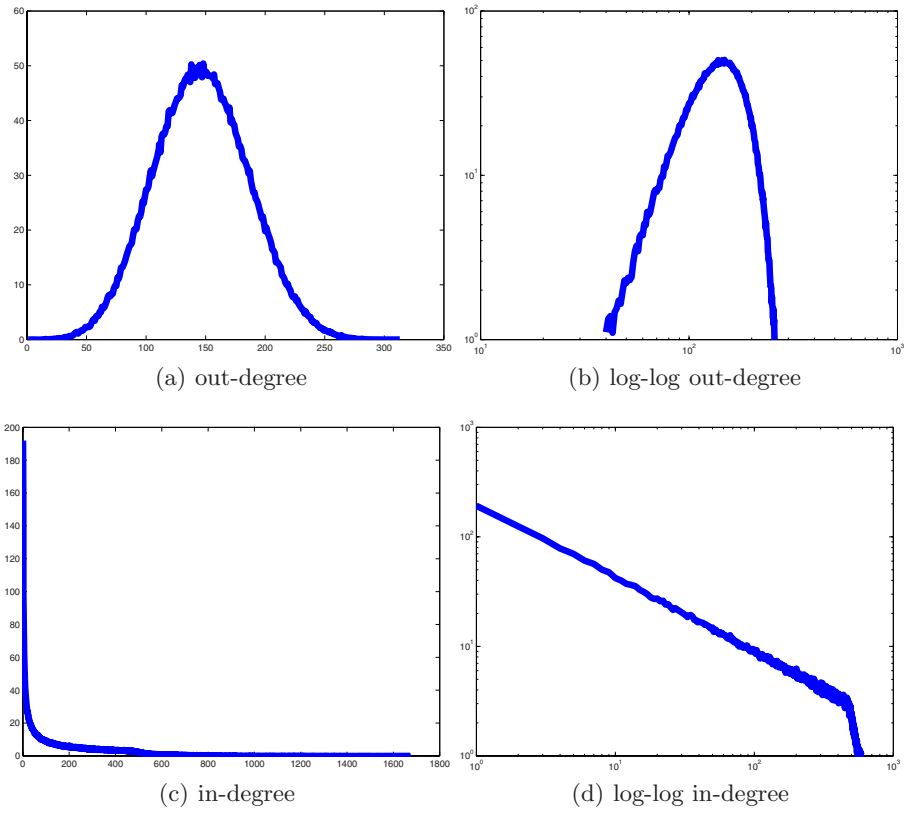

Fig. 1. Average degree sequences for given directed random dot product graph

of the underling space [11], it is natural to consider whether the random dot product graphs exhibit similar behavior. In this section we characterize the conductance of the geometric cuts in the undirected random dot product graph model. Specifically, we show that any bad cuts have no semantic content. That is, if low conductance cuts exist they are essentially non-geometric.

Before discussing the nature of the geometric cuts in $G(\mathbf{W}, n)$ we first need some preliminary definitions. For any region R we will abuse notation and refer to the set of vertices whose vectors are contained in the region as R. We will also, for notational convenience, denote by W_R the expectation of $\mathbf{W}$ restricted to R for any $\mathbf{W}$-measurable set R. Using notation standard from conductance we will denote by $\mathrm{Vol}(R)$ the sum of the degrees for all vertices in R and by $C(R, \overline{R})$ the number of edges crossing the cut $(R, \overline{R})$. Finally, we also denote by $\mathbb{P}(R)$ the probability that a random variable distributed as $\mathbf{W}$ lies within the region R. The conductance of the cut $(R, \overline{R})$, $\Phi_R(G(\mathbf{W}, n))$, is defined as $C(R, \overline{R}) / \min\left\{\mathrm{Vol}\, R, \mathrm{Vol}\, \overline{R}\right\}$. With these definitions and a multidimensional generalization of the Chernoff Bound [22], we have the following results on the nature of geometric cuts in $G(\mathbf{W}, n)$.

Theorem 3. *Let R be a fixed subset of $\mathbb{R}^d$ and let $\mathbf{W}$ be a distribution on $\mathbb{R}^d$ that satisfies the inner product condition. Then almost surely*

$$\lim_{n\to\infty} \Phi_R(G(\mathbf{W}, n)) \geq \frac{\mathbb{P}\left(\overline{R}\right) \langle W_R, W_{\overline{R}} \rangle}{\langle W_R, \mathbb{E}\left[\mathbf{W}\right] \rangle}, \tag{9}$$

when $\mathbb{P}\left(R\right) \|W_R\|^2 \leq \mathbb{P}\left(\overline{R}\right) \|W_{\overline{R}}\|^2$.

This results establishes that any fixed region does not induce a bad cut. However it leaves open the possibility that there is some sequence of regions giving arbitrarily small conductance. That is, it may be possible that for an arbitrary $c > 0$ there is some region R_c such that the conductance induced by the cut $(R_c, \overline{R}_c)$ is constant but less than c. In fact, by using the inner product condition we may show the following result:

Theorem 4. *For a fixed distribution $\mathbf{W}$ satisfying the inner product condition, $\inf_R \lim_{n\to\infty} \Phi_R(G(\mathbf{W}, n))$ is bounded below, where the infimum is taken over $\mathbf{W}$-measurable sets R.*

By combining the results of Theorem 3 and Theorem 4, we conclude that if $\mathbf{W}$ satisfies the inner product condition there is some $\alpha > 0$, depending only on $\mathbf{W}$, such that for any region $R \subseteq \mathbb{R}^d$, $\lim_{n\to\infty} \Phi_R(G(\mathbf{W}, n)) > \alpha$, with high probability. Specifically, any fixed partition $(R, \overline{R})$ of $\mathbb{R}^d$ has constant conductance independent of $(R, \overline{R})$. Thus, in contrast to the work of Flaxman, Frieze and Vera, where they showed that the geometric preferential attachment model has bad cuts induced entirely by the geometry, if the random dot product graph model has bad cuts they are entirely non-geometric. This does, however, leave open the question of what happens for non-fixed geometric regions and non-geometric partitions. Although we believe that the conductance of the random dot product graph model is asymptotically constant, the slow rate of convergence of this result leaves open the possibility that for every n there is a positive probability that some region has conductance smaller than α. Furthermore, since this result is inherently geometric, it says little about the case where $\mathbf{W}$ is not a continuous distribution. For instance, if $\mathbf{W}$ contains a point mass, then there is no way to geometrically place vertices generated by a point mass on opposing sides of a partition, whereas a partition of the vertices can clearly separate those vertices. Thus, fully resolving the conductance of the random dot product graph model will require a fundamental non-geometric insight into the structure of these graphs as in the work of Mihail, et al. [23].

7 Future Work

There are some natural questions that this work brings up. Perhaps the most pressing is the development of a sparse, or preferably, a variable density analogue of both $G(\mathbf{W}, n)$ and $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$. Although, as we noted above, the presence of $\Omega\left(n^2\right)$ edges is not as major an objection as it once was for social networks, it

still limits the models' general applicability. Thus a natural sparse generalization would broaden the applicability of these models. We have positive results in a preliminary work in this direction [24].

Also, given that the result that reinvigorated the study of social networks was Milgram's experimental result on the navigability of the "real world" social network, [25,26], it is reasonable to consider under what conditions short paths can be found in $G(\mathbf{W}, n)$. Kleinberg's result on the navigation of the grid with power-law shortcuts showed that navigation is sensitive to parameters of the model [27], however we feel that the additional semantic information in $G(\mathbf{W}, n)$ will allow navigation under more general conditions.

Finally, from a simulation point of view, it would be desirable to have a means of rapidly generating samples from $G(\mathbf{W}, n)$ or $\overrightarrow{G}(\mathbf{X}, \mathbf{Y}, n)$. Kraetzl, Nickel, and Scheinerman [1] discuss a thresholding modification of the natural means of generation that will produce an approximate sample. However, it is not immediately obvious how much the loss of edges due to thresholding will affect any given property of the sample graph. Thus, for serious simulation purposes, some means of estimating the effect of the thresholding or a clever way of reducing the overall computation time would seem to be necessary.

Acknowledgments. The first author is grateful to Milena Mihail for her advice and support in the preparation of this document. The authors would also like to thank the anonymous referees for the helpful comments.

References

1. Kraetzl, M., Nickel, C., Scheinerman, E.R.: Random dot product graphs: A model for social netowrks. Preliminary Manuscript (2005)
2. Kraetzl, M., Nickel, C., Scheinerman, E.R., Tucker, K.: Random dot product graphs (July 2005), `http://www.ipam.ucla.edu/abstract.aspx?tid=5498`
3. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Modern Phys. 74(1), 47–97 (2002)
4. Achlioptas, D., Kempe, D., Clasuet, A., Moore, C.: On the bias of traceroute sampling or, power-law degree distributions in regular graphs. In: STOC 2005. Proc. of the 37$^{\text{th}}$ ACM Symposium on the Theory of Computer Science (2005)
5. Lakhina, A., Byers, J.W., Crovella, M., Xie, P.: Sampling biases in IP topology measurements. In: INFOCOM 2003. 22$^{\text{nd}}$ Joint Conference of the IEEE Computer and Communications Societies (2003)
6. Durrett, R.: Random graph dynamics. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge (2007)
7. Chung, F., Galas, D.J., Dewey, T.G., Lu, L.: Duplication models for biological networks. Journal of Computational Biology (2003)
8. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tompkins, A., Upfal, E.: The web as a graph. In: PODS 2000. Proc. of the 19$^{\text{th}}$ ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 1–10. ACM Press, New York (2000)
9. Bornholdt, S., Schuster, H.G. (eds.): Handbook of graphs and networks. From the genome to the internet. Wiley-VCH, Weinheim (2003)

10. Newman, M.E.J.: Assortative mixing in networks. Physical Review Letters 89 (2002)
11. Flaxman, A.D., Frieze, A.M., Vera, J.: A geometric preferential attachment model of networks. Internet Math. 3(2), 187–205 (2006)
12. Caldarelli, G., Capocci, A., de Los Rios, P., Muñoz, M.A.: Scale-Free Networks from Varying Vertex Intrinsic Fitness. Physical Review Letters 89(25) (2002)
13. Azar, Y., Fiat, A., Karlin, A., McSherry, F., Saia, J.: Spectral analysis of data. In: STOC 2001. Proc. of the 33rd ACM Symposium on Theory of Computing, pp. 619–626. ACM Press, New York (2001)
14. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM Trans. Knowl. Discov. Data 1(1) (2007)
15. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences 102(33), 11623–1162 (2005)
16. Bollobás, B.: Modern graph theory. In: Bollobás, B. (ed.) Graduate Texts in Mathematics, vol. 184, Springer, New York (1998)
17. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization; Analysis, Algorithms, and Engineering Applications, SIAM, Philadelphia, PA (2001)
18. Hörmnn, W., Leydold, J.: Random-number and random-variate generation: automatic random variate generation for simulation input. In: Winter Simulation Conference, pp. 675–682 (2000)
19. Scheinerman, E.R., Tucker, K.: Exact and asymptotic dot product representations of graphs i: Fundamentals (Submitted, 2007)
20. Scheinerman, E.R., Tucker, K.: Exact and asymptotic dot product representations of graphs ii: Characterization and recognition (Submitted, 2007)
21. Scheinerman, E.R., Tucker, K.: Modelling graphs using dot product representations. (preparation, 2007)
22. Alon, N., Spencer, J.H.: The Probabilistic Method. In: Wiley-Interscience Series in Discrete Mathematics and Optimization, 2nd edn., Wiley-Interscience, New York (2000)
23. Mihail, M., Papadimitriou, C., Saberi, A.: On certain connectivity properties of the internet topology. J. Comput. System Sci. 72(2), 239–251 (2006) (FOCS 2003 Special Issue)
24. Young, S.J.: Sparse random dot product graphs. (preparation, 2007)
25. Milgram, S.: The small world problem. Psychology Today (1967)
26. Milgram, S., Travers, J.: An experimental study of the small world problem. Sociometry 32(4), 425–443 (1969)
27. Kleinberg, J.M.: The small world phenomenon: an algorithmic perspective. In: STOC 1999. Proc. of the 32nd ACM Symposium on the Theory of Computer Science (1999)