Structural Properties of Random Graph Models

András Faragó

Department of Computer Science
Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
P.O. Box 830688, MS-EC31
Richardson, Texas 75083-0688, U.S.A.
E-mail: farago@utdallas.edu

Abstract

Many different random graph constructions are used to model large real life graphs. Often it is not clear, however, how the strength of the different models compare to each other, e.g., when does it hold that a certain model class contains another. We are particularly interested in random graph models that arise via abstract geometric constructions, motivated by the fact that these graphs can model certain wireless communication networks. We set up a general framework to compare the strength of random graph models, and present some results about the equality, inequality and proper containment of certain model classes, as well as some open problems.

1 Introduction

Large real life graphs are often modeled by various random graph contructions, see, e.g. Bornholdt & Shuster (2003), Franceschetti & Meester (2007), Penrose (2003) and many further references therein. In many cases it is not at all clear how the modeling strength of differently generated random graph model classes relate to each other. We would like to initiate a systematic investigation of such issues. Our approach was originally motivated to capture properties of the random network topology of wireless communication networks. We started some investigations in Faragó (2007), but here we elevate it to a more abstract level that makes it possible to compare the strength of different classes of random graph models.

Supported in part by NSF Grant CCF-0634848.

2 Classes of Random Graph Models

2.1 General Random Graph Models

Let us first explain what we mean by random graphs and a random graph model in the most general sense. In full generality, by a *random graph* on a fixed number of vertices (n) we mean a random variable that takes its values in the set of all undirected graphs[1] on n vertices[2]. We denote a random graph on n nodes by G_n. At this point, it is still completely general, it can be generated by any mechanism, with arbitrary dependencies among its parts, it is just *any* graph-valued random variable, taking its values among undirected graphs on n nodes.

Definition 1 (General random graph model) *A* random graph model *is given by a sequence of graph valued random variables, one for each possible value of n:*

$$\mathcal{M} = (G_n;\ n \in \mathbf{N}).$$

The set of all such models is denoted by **GEN**.

It is worth noting that even though the model is defined by random graphs for fixed values of n, when we apply it to actually generate a random graph, we do not have to use a fixed number of nodes. For example, let ν be a Poisson random variable. We can then consider the random graph G_ν which is generated such that we first draw the random value of ν and then take ν^{th} entry from the random graph sequence $\mathcal{M} = (G_1, G_2, \ldots)$. Thus, the random graph model $\mathcal{M}$ serves as a defining basis, from which we can generate random graphs in different ways. The simplest way, of course, is just to take G_n for a fixed n, but, as the example shows, it is not the only way.

2.2 Geometric Random Graph Models

Let us now introduce a model class that reflects a typical feature of geometric random graph models. This feature is that in geometric random graphs the primary random choice is picking random nodes from some domain and then the edges are already determined by some geometric property (typically some kind of distance) of the random nodes. We elevate this approach to an abstract level that, as will be shown later, actually turns out to be no less general than the totally unrestricted model. Our model is built of the following components:

[1] The approach could also be developed for directed graphs in a similar way, but in this paper we stay with undirected graphs.

[2] We use the words "vertex" and "node" interchangeably.

- **Node variables.** The nodes are represented by an infinite sequence $X_1, X_2, \ldots$ of random variables, called *node variables*. They take their values in an arbitrary (nonempty) set S, which is called the *domain* of the model. When a random graph on n nodes is generated, then we use the first n entries of the sequence, that is, $X_1, \ldots, X_n$ represent the nodes in G_n. It is important to note that we do not require the node variables to be independent.

- **Edge functions.** We denote by $Y_{ij}^{(n)} \in \{0,1\}$ the indicator of the edge between nodes X_i, X_j in the random graph G_n. Since loops are not allowed, we always assume $i \neq j$, without repeating this condition each time. The (abstract) geometric nature of the model is expressed by the requirement that the random variables $Y_{ij}^{(n)}$ are determined by the nodes $X_1, \ldots, X_n$, possibly with additional independent randomization. Specifically, we assume that there exist functions $f_{ij}^{(n)}$, such that

$$Y_{ij}^{(n)} = f_{ij}^{(n)}(X_1, \ldots, X_n, \xi_{ij})$$

  where ξ_{ij} is a random variable that is uniformly distributed on $[0,1]$ and is independent of all the other defining random variables of the model (i.e, the node variables and all the other ξ_{kl} variables). Henceforth the role of ξ_{ij} is referred to as *independent randomization*[3]. The undirected nature of the graph is expressed by the requirement $Y_{ij}^{(n)} = Y_{ji}^{(n)}$, which can simply be enforced by computing all values for $i < j$ only and defining the $i > j$ case by exchanging i and j.

We use the following **notational convention:** whenever a function is distinguished by certain parameters within some family of functions, such as $f_{ij}^{(n)}$ above, then it is assumed that the function "knows" its own parameters. In other words, the parameter values can be used in the definition of the function. Conversely, whatever information is used in computing the function should occur either as a variable or an explicitely shown parameter.

Definition 2 (Abstract geometric model) *The class of all models that have the structure explained above is called* **GEOM**.

A model $\mathcal{M} \in$ **GEOM**, no matter how general it can be, still has a restricted structure. Therefore, one may ask whether *every* model in **GEN** can be represented in such a way. To make it precise when two models or model classes are considered equivalent, let us introduce the following definition.

Definition 3 (Equivalence) *Two random graph models $\mathcal{M} = (G_n;\ n \in \mathbf{N})$ and $\widetilde{\mathcal{M}} = (\widetilde{G}_n;\ n \in \mathbf{N})$ are called* equivalent*, denoted by $\mathcal{M} \sim \widetilde{\mathcal{M}}$, if for any graph G*

$$\Pr(G_n = G) = \Pr(\widetilde{G}_n = G)$$

holds, where equality of graphs means that they are isomorphic.

[3]Note that the specified distribution of ξ_{ij} does not impose a restriction, since the functions $f_{ij}^{(n)}$ are arbitrary.

Definition 4 (Containment, equivalence and disjointness of model classes) *Let $\mathbf{C_1}, \mathbf{C_2}$ be two classes of random graph models. We say that $\mathbf{C_2}$ contains $\mathbf{C_1}$, denoted by $\mathbf{C_1} \preceq \mathbf{C_2}$, if for every $\mathcal{M}_1 \in \mathbf{C_1}$ there is an $\mathcal{M}_2 \in \mathbf{C_2}$, such that $\mathcal{M}_1 \sim \mathcal{M}_2$. If $\mathbf{C_1} \preceq \mathbf{C_2}$ and $\mathbf{C_2} \preceq \mathbf{C_1}$ both hold, then the two classes are called* equivalent*, denoted by $\mathbf{C_1} \simeq \mathbf{C_2}$. If there exist no models $\mathcal{M}_1 \in \mathbf{C_1}$ and $\mathcal{M}_2 \in \mathbf{C_2}$ with $\mathcal{M}_1 \sim \mathcal{M}_2$, then the two classes are called* disjoint*, denoted by $\mathbf{C_1} \cap \mathbf{C_2} = \emptyset$.*

Now we may ask whether **GEOM** $\simeq$ **GEN** holds or not. We show later that it does, even with more restrictions on **GEOM**. To this end, we introduce some restricting conditions to the model class **GEOM**. As a simple notation, whenever some restrictions $R_1, \ldots, R_k$ are applied, the arising class is denoted by $\mathbf{GEOM}(R_1, \ldots, R_k)$.

2.3 Subclasses of GEOM

The first considered restriction is called *locality*. Up to now we allowed that an edge in G_n can depend on all the nodes, and the dependence expressed by the $f_{ij}^{(n)}$ functions can be arbitrary and different for each edge. To get a little closer to the usual geometric random graph model, we introduce the condition of locality. Informally, it restricts the dependence of an edge to its endpoints, in a homogeneous way, but still via an *arbitary* function.

Definition 5 (Locality) *A model $\mathcal{M} \in$ **GEOM** is called* local*, if for every n and $i, j \leq n$ the existence of an edge between X_i, X_j depends only on these nodes. Moreover, the dependence is the same for every i, j, possibly with independent randomization. That is, there are functions $f^{(n)}$ such that the edge indicators are expressible as*

$$Y_{ij}^{(n)} = f^{(n)}(X_i, X_j, \xi_{ij})$$

where ξ_{ij} represents the independent randomization. The set of local models in **GEOM** *is denoted by* $\mathbf{GEOM}(loc)$.

Note: with our notational convention $f^{(n)}$ can depend on its variables and on n. On the other hand, it has no access to the value of i and j, unless they are somehow contained in X_i, X_j, in a way that makes it possible to extract them without using anything else than the explicitely listed information.

Another restriction that we consider is a condition on the distribution of the vertices. To introduce it, let us first recall a concept from probability theory, called exchangeability.

Definition 6 (Exchangeable random variables) *A finite sequence $\xi_1, \ldots, \xi_n$ of random variables is called* exchangeable *if for any permutation σ of $\{1, \ldots, n\}$, the joint distribution of $\xi_1, \ldots, \xi_n$ is the same as the joint distribution of $\xi_{\sigma(1)}, \ldots, \xi_{\sigma(n)}$. An infinite sequence of random variables is called exchangeable if every finite initial segment of the sequence is exchangeable.*

Exchangeability can be equivalently defined such that taking any $k \geq 1$ of the random variables, say, $\xi_{j_1}, \ldots, \xi_{j_k}$, their joint distribution does not depend

on which particular k of them are taken. Note that independent, identically distributed (i.i.d.) random variables are always exchangeable, but the converse is not true, so this is a larger family.

Now let us introduce the condition that we use to restrict the arbitrary dependence of node variables.

Definition 7 (Name invariance) *A random graph model $\mathcal{M} \in$ **GEOM** is called* name invariant*, if its node variables are exchangeable. The class of such models is denoted by* $\mathbf{GEOM}(inv)$.

We call it the *name invariance* of the model because it means the names (the indices) of the nodes are irrelevant in the sense that the joint probabilistic behavior of any fixed number of nodes is invariant to renaming (reindexing) the nodes. In particular, it also implies that each single node variable X_i has the same probability distribution (but they do not have to be independent).

A simple example for a dependent, yet still name invariant, node generation process is a "clustered uniform" node generation. As an example, let S be a a sphere in 3-dimensional space, i.e., the surface of a 3-dimensional ball. Let R be the radius of the ball. Let us first generate a pivot point Y uniformly at random from S. Then generate the nodes $X_1, X_2, \ldots$ uniformly at random and independently of each other from the neighborhood of radius $r \ll R$ of the random pivot point Y (within the sphere). It is directly implied by the construction that exhangeability holds. Moreover, any particular X_i will be uniformly distributed over the *entire* sphere, since Y is uniform over the sphere. On the other hand, the X_i are far from independent of each other, since they cluster around Y, forcing any two of them to be within distance $2r$. The example can be generalized to applying several pivot points and non-uniform distributions, creating a more sophisticated clustering.

It is worth mentioning that *any* finite sequence $X_1, \ldots, X_n$ of random variables can be easily transformed into an exchangeable sequence by taking a *random* permutation σ of $\{1, \ldots, n\}$ and defining the transformed sequence by $\widetilde{X}_i = X_{\sigma(i)}$. The resulting joint distribution will be

$$\Pr(\widetilde{X}_1 = x_1, \ldots, \widetilde{X}_n = x_n) =$$

$$\frac{1}{n!} \sum_{\sigma} \Pr(X_{\sigma(1)} = x_1, \ldots, X_{\sigma(n)} = x_n)$$

where σ in the summation runs over all possible permutations of $\{1, \ldots, n\}$. Even though this simple construction does not work for infinite sequences, in many practically relevant cases there is vanishing difference between a very long finite and an actually infinite sequence.

A stronger restriction is if we want the node variables to be independent, not just exchangeable.

Definition 8 (Free geometric model) *A random graph model $\mathcal{M} \in$ **GEOM** is called* free*, if its node variables are mutually independent. The class of such models is denoted by* $\mathbf{GEOM}(free)$.

2.4 Other Model Classes

We define some other classes of random graph models, relating to some properties that are important in the applications of these models.

Definition 9 (Bounded expected degree model) *A random graph model $\mathcal{M} \in$ **GEN** is called a* bounded expected degree model *if there exists a constant C such that*

$$\overline{d}(n) = \frac{2\mathrm{E}(e(G_n))}{n} \leq C$$

for every n, where $e(G_n)$ denotes the number of edges in G_n. The class of bounded expected degree models is denoted by **BD**.

Since $2e(G_n)/n$ is the average degree in G_n, therefore, $\overline{d}(n) = 2\mathrm{E}(e(G_n))/n$ is the expected average degree. Often the expected degree of each individual node is also equal to $\overline{d}(n)$, but in a general model it may not hold. Note that even if the expected degree of each node is equal to the expected average degree, it does not mean that the actual (random) degrees are also equal, so G_n may be far from regular.

Another important property of random graph models is asymptotically almost sure (a.a.s.) connectivity.

Definition 10 (Connected model) *A random graph model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in$ **GEN** is called* connected *if*

$$\lim_{n\to\infty} \Pr(G_n \text{ is connected}) = 1.$$

The class of connected models is denoted by **CONN**.

Note: Whenever we write down a limit, such as the one above, we also assume that the limit exists.

Often the requirement of full connectivity is too strong, so we define a relaxed version of it and the corresponding model class.

Definition 11 (β-connectivity) *For a real number $0 \leq \beta \leq 1$, a graph G on n vertices is called β-connected if G contains a connected component on at least βn nodes.*

When we consider a sequence of graphs with different values of n, then the parameter β may depend on n. When this is the case, we write β_n-connectivity. Note that even if $\beta_n \to 1$, this is still weaker then full connectivity in the limit. For example, if $\beta_n = 1 - 1/\sqrt{n}$, then we have $\beta_n \to 1$, but there can be still $n - \beta_n n = \sqrt{n}$ nodes that are not part of the largest connected component.

Definition 12 (β_n-connected model) *A random graph model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in$ **GEN** is called* β_n-connected *if*

$$\lim_{n\to\infty} \Pr(G_n \text{ is } \beta_n\text{-connected}) = 1.$$

The class of β_n-connected models is denoted by β_n-**CONN**.

It is clear from the definitions that with $\beta_n \equiv 1$, the class 1-**CONN** is the same as **CONN**. But if we only know that $\beta_n \to 1$, then β_n-**CONN** becomes a larger class.

Finally, let us define some classes that restrict the indepedence structure of the edges. Let e be a (potential) edge. We regard it as a 0-1 valued random

variable, indicating whether the edge is in the random graph or not. The probability that an edge e exists is $\Pr(e = 1)$, but we simply denote it by $\Pr(e)$. We similarly write $\Pr(e_1, \ldots, e_k)$ instead of $\Pr(e_1 = 1, \ldots, e_k = 1)$.

Definition 13 (Independent disjoint edges) *A random graph model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in \mathbf{GEN}$ is said to have* independent disjoint edges *if any set $e_1 \ldots, e_k$ of pairwise disjoint edges are independent as random variables. That is,*

$$\Pr(e_1, \ldots, e_k) = \Pr(e_1) \ldots \Pr(e_k)$$

holds whenever $e_1, \ldots, e_k$ are pairwise disjoint. The class of models with independent disjoint edges is denoted by **IDE**.

Definition 14 (Positively correlated edges) *A random graph model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in \mathbf{GEN}$ is said to have* positively correlated edges *if any set $e_1 \ldots, e_k$ of distinct edges are positively correlated in the sense of*

$$\Pr(e_1, \ldots, e_k) \geq \Pr(e_1) \ldots \Pr(e_k).$$

The class of models with positively correlated edges is denoted by **POS**.

3 Results

Let us first address the question how the various restrictions influence the modeling strength of **GEOM**. The motivation is that one might think that a concept like locality imposes a significant restriction on the model. After all, it severely restricts which node variables can directly influence the existence of an edge. For example, it seems to exclude situations when the existence of an edge between X_i and X_j is based on whether one of them is among the k nearest neighbors of the other, according to some distance function (often called k-nearest neighbor graph).

Surprisingly, it turns out that locality alone does not impose *any* restriction at all on the generality of the model. Not just any model in **GEOM** can be expressed by a local one, but this remains true even if we want to express an *arbitrary* random graph model in **GEN**.

Theorem 1 *Let $\widetilde{\mathcal{M}} = (\widetilde{G}_n;\ n \in \mathbf{N}) \in \mathbf{GEN}$ be an arbitrary random graph model. Then there exists a another model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in \mathbf{GEOM}(loc)$ such that $\mathcal{M} \sim \widetilde{\mathcal{M}}$.*

Proof. Let $\widetilde{Y}_{ij}^{(n)}$ denote the edge indicators in $\widetilde{\mathcal{M}}$. We show that a $\mathcal{M} \in \mathbf{GEOM}(loc)$ can be chosen such that its edge indicators $Y_{ij}^{(n)}$ satisfy $Y_{ij}^{(n)} = \widetilde{Y}_{ij}^{(n)}$, which implies that the two models are equivalent.

Let Q be the set of all 0-1 matrices of all possible finite dimensions. For the domain S of $\mathcal{M}$ we choose the set of all infinite sequences with entries in Q. Let us define the node variable X_i such that

$$X_i = (Z_i^{(1)}, Z_i^{(2)}, \ldots)$$

where $Z_i^{(n)}$ is an $(n+1) \times n$ sized 0-1 matrix with entries $Z_i^{(n)}[k, \ell] = \widetilde{Y}_{k,\ell}^{(n)}$ for $k \neq \ell$ and $k, \ell \leq n$, $Z_i^{(n)}[k, k] = 0$ and the last row $Z_i^{(n)}[n+1, .]$ contains the binary encoding of i. Then the edge functions for $\mathcal{M}$ can be defined as

$$f^{(n)}(X_i, X_j, \xi_{ij}) = Z_i^{(n)}[i, j].$$

This indeed defines $f^{(n)}$, since knowing n the matrix $Z_i^{(n)}$ can be obtained as the n^{th} component of X_i. The value of i can be read out from the last row of $Z_i^{(n)}$. Similarly, the value of j can be read out from the last row of $Z_j^{(n)}$, which is the n^{th} component of X_j. Then the value of $Z_i^{(n)}[i, j]$ can be looked up. (The functions do not use the independent randomization). This definition directly implies that $\mathcal{M}$ is local, as $f^{(n)}$ does not use node variables other than X_i, X_j and the same function applies to any pair of nodes. Furthermore,

$$Y_{ij}^{(n)} = f^{(n)}(X_i, X_j, \xi_{ij}) = Z_i^{(n)}[i, j] = \widetilde{Y}_{ij}^{(n)}$$

holds, completing the proof.

♠

Next we show that a similar result holds for the restriction of name invariance.

Theorem 2 *Let $\widetilde{\mathcal{M}} = (\widetilde{G}_n;\ n \in \mathbf{N}) \in \mathbf{GEN}$ be an arbitrary random graph model. Then there exists a another model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in \mathbf{GEOM}(inv)$ such that $\mathcal{M} \sim \widetilde{\mathcal{M}}$.*

Proof. We show that the name invariant model $\mathcal{M} \in \mathbf{GEOM}(inv)$ can be chosen such that its edge indicators $Y_{ij}^{(n)}$ satisfy $Y_{ij}^{(n)} = \widetilde{Y}_{ij}^{(n)}$, where the $\widetilde{Y}_{ij}^{(n)}$ denote the edge indicators in $\widetilde{\mathcal{M}}$.

Let $Z_n = [\widetilde{Y}_{ij}^{(n)}]$ be an $n \times n$ matrix, containing all edge indicators of $\widetilde{G}_n$. Define X_i as an infinite sequence

$$X_i = (Z_1, Z_2, \ldots).$$

Since X_i is defined without using the value of i, we have that all the X_i are equal, which is a trivial case of name invariance. (All random node variables being equal, re-indexing clearly cannot change anything.) Then, following the edge function format of **GEOM**, we can define the edge functions by

$$f_{ij}^{(n)}(X_1, \ldots, X_n, \xi_{ij}) = Z_n[i, j].$$

(The independent randomization is not used.) This edge function is well defined, since, knowing n, the array Z_n can be read out from any of the X_i and in the general **GEOM** model the functions can directly depend on i and j. As, by definition, $Z_n[i, j] = \widetilde{Y}_{ij}^{(n)}$, we obtain

$$Y_{ij}^{(n)} = f_{ij}^{(n)}(X_1, \ldots, X_n, \xi_{ij}) = Z_n[i, j] = \widetilde{Y}_{ij}^{(n)}$$

which completes the proof.

♠

Since we know by definition $\mathbf{GEOM}(loc) \preceq \mathbf{GEOM}$ and $\mathbf{GEOM}(inv) \preceq \mathbf{GEOM}$, as well as $\mathbf{GEOM} \preceq \mathbf{GEN}$, the theorems immediately imply the following corollary.

Corollary 1 $\mathbf{GEOM}(loc) \simeq \mathbf{GEOM}(inv) \simeq \mathbf{GEOM} \simeq \mathbf{GEN}$.

We have seen above that neither locality nor name invariance can restrict full generality. Both restrictions, if applied alone, still allow that an *arbitrary* random graph model is generated. This situation naturally leads to the question: what happens if the two restrictions are applied *together?* At first, one might think about it this way: if the set of local models and the set of name invariant models are both equal to the set of general models, then their intersection should also be the same. This would mean that even those models that are both local and name invariant are still fully general.

The above argument, however, is not correct. Although we know from Corollary 1 that $\mathbf{GEOM}(loc) \cap \mathbf{GEOM}(inv) \simeq \mathbf{GEN}$, but it does not imply that $\mathbf{GEOM}(loc, inv) \simeq \mathbf{GEOM}(loc) \cap \mathbf{GEOM}(inv)$ also holds. In fact, the latter does not hold, which will be obtained as a consequence of the following theorem. The theorem proves the interesting thing that joint locality and name invariance makes it impossible that a model satisfies bounded expected degree and (almost) connectivity at the same time.

Theorem 3 *Let $\beta_n \to 1$ be a sequence of positive reals. Then*

$$\mathbf{BD} \cap \beta_n\text{–}\mathbf{CONN} \cap \mathbf{GEOM}(loc, inv) = \emptyset$$

holds.

Proof. Consider a model $\mathcal{M} = (G_n;\ n \in \mathbf{N}) \in \mathbf{GEOM}(loc, inv)$. Let I_n denote the (random) number of isolated nodes in G_n. First we show that

$$\mathrm{E}(I_n) \geq n \left(1 - \frac{\overline{d}(n)}{n-1}\right)^{n-1} \tag{1}$$

holds[4]. Note that since our model is abstract and does not involve any real geometry, one has to be careful to avoid using such intuition that may appeal geometrically, but does not follow from the abstract model.

First, observe the following: name invariance implies that for any function g of the node variables and for any permutation σ of $\{1, \ldots, n\}$ we have

$$\mathrm{E}(g(X_1, \ldots, X_n)) = \mathrm{E}(g(X_{\sigma(1)}, \ldots, X_{\sigma(n)})).$$

Since the probability that a particular node has any given degree k is also expressible by such a function, therefore, the probability distribution of the node degree must be the same for all nodes (but the degrees, as random variables, may not be independent). As a consequence, the expected degree of each node is the same, which then must be equal to the expected average degree $\overline{d}(G_n)$.

Let us pick a node X_i. We derive a lower bound on the probability that X_i is isolated, i.e., its degree is 0. Due to the above symmetry considerations, it does not matter which node is chosen, so we can take $i = 1$. Let $\mathcal{I}_n$ be the (random) set of isolated nodes in G_n. What we want to compute is a lower bound on $\Pr(X_1 \in \mathcal{I}_n)$. Then we are going to use the fact that

$$\mathrm{E}(I_n) = \mathrm{E}(|\mathcal{I}_n|) = \sum_{i=1}^{n} \Pr(X_i \in \mathcal{I}_n)$$

Note that, due to the linearity of expectation, this remains true even if the events $\{X_i \in \mathcal{I}_n\}$ are not independent, which is typically the case. Then, by the symmetry considerations, we can utilize that $\Pr(X_i \in \mathcal{I}_n)$ is independent of i, yielding $\mathrm{E}(I_n) = n \Pr(X_1 \in \mathcal{I}_n)$.

In order to derive a lower bound on $\Pr(X_1 \in \mathcal{I}_n)$, we need a fundamental result from probability theory, called *de Finetti's Theorem*[5] This theorem says that if an infinite sequence $\xi_1, \xi_2, \ldots$ of 0-1 valued random variables[6] is exchangeable, then the following hold:

(i) The limit

$$\eta = \lim_{N \to \infty} \frac{\xi_1 + \ldots + \xi_N}{N} \tag{2}$$

exists[7] with probability 1.

(ii) For any N and for any system $a_1, \ldots, a_N \in \{0, 1\}$ of outcomes with $s = \sum_{i=1}^{N} a_i$

$$\Pr(\xi_1 = a_1, \ldots, \xi_N = a_N) =$$

$$= \int_0^1 x^s (1-x)^{N-s} dF_\eta(x)$$

holds, where F_η is the probability distribution function of η.

(iii) The ξ_i are conditionally independent and identically distributed (conditionally i.i.d.), given η, that is,

$$\Pr(\xi_1 = a_1, \ldots, \xi_N = a_n \,|\, \eta) = \prod_{i=1}^{N} \Pr(\xi_i = a_i \,|\, \eta).$$

Informally, de Finetti's theorem says that exchangeable 0-1 valued random variables, even if they are not independent, can always be represented as a mixture of Bernoulli systems of random variables. It is important to note, however, that even though the statements (ii) and (iii) refer to finite initial segments of the sequence $\xi_1, \xi_2, \ldots$, it is necessary that the entire *infinite* sequence is exchangeable. For finite sequences the theorem may not hold, counterexamples are known for the finite case Stoyanov (1987).

Let us now define the infinite sequence of 0-1 valued random variables

$$e_j = f^{(n)}(X_1, X_j, \xi_{1j}), \qquad j = 2, 3 \ldots$$

[4] It is worth noting that even when $\mathrm{E}(I_n) \to \infty$ is the case, this fact alone may not *a priori* preclude the possibility of a.a.s. β_n-connectivity, even with $\beta_n \equiv 1$. For example, if G_n is connected with probability $1 - 1/\sqrt{n}$ and consists of n isolated nodes with probability $1/\sqrt{n}$, then $\mathrm{E}(I_n) = n/\sqrt{n} \to \infty$, but $\Pr(G_n \text{ is connected}) = 1 - 1/\sqrt{n} \to 1$.

[5] It was first published in de Finetti (2003). Being a classical result, it can be found in many advanced textbooks on probability.

[6] Various extensions exist to more general cases, see, e.g., Kallenberg (2005), but for our purposes the simplest 0-1 valued case is sufficient.

[7] Note that exhangeability implies that all ξ_i have the same expected value, so in case they were independent, then the strong law of large numbers would apply and the limit would be the common expected value, with probability 1. Since, however, the ξ_i are not assumed independent (only exchangeable), therefore, the average may not tend to a constant, it can be a non-constant random variable in $[0, 1]$.

Of these, $e_2, \ldots, e_n$ are the indicators of the edges with one endpoint at X_1. But the function $f^{(n)}$ is defined for any $(x, y, z) \in S \times S \times [0, 1]$, so nothing prevents us to define the *infinite* sequence e_j; $j = 2, 3, \ldots$, by taking more independent and uniform $\xi_{1j} \in [0, 1]$ random variables.

Observe now that the sequence e_j; $j = 2, 3, \ldots$ is an infinite exchangeable sequence of 0-1 valued random variables. Only the exchangeability needs proof. If we take any k indices $j_1, \ldots, j_k$, then the joint distribution of $e_{j_1}, \ldots, e_{j_k}$ depends only the joint distribution of $X_{j_1}, \ldots, X_{j_k}$, plus the independent randomization. If we replace $j_1, \ldots, j_k$ by other k indices, then it will not change the joint distribution of the k node variables, due to their assumed exhangeability. The independent randomization also does not change the joint distribution, since the ξ_{1j} are i.i.d, so it does not matter which k are taken. Furthermore, the locality of the model implies that each e_j depends on one X_j (besides X_1) so taking another k cannot change how many node variables will any subset of the e_j share. Thus, for any k, the joint distribution of $e_{j_1}, \ldots, e_{j_k}$ does not depend on which k indices are chosen, proving that e_j; $j = 2, 3, \ldots$ is an infinite exchangeable sequence of 0-1 valued random variables.

Now, by de Finetti's Theorem, there is a random variable $\eta \in [0, 1]$, such that the e_j are conditionally i.i.d, given η. Then we can write

$$\begin{aligned}
\Pr(X_1 \in \mathcal{I}_n) &= \Pr(e_2 = \ldots = e_n = 0) \\
&= \mathrm{E}(\Pr(e_2 = \ldots = e_n = 0 \,|\, \eta)) \\
&= \mathrm{E}\left(\prod_{j=2}^{n} (\Pr(e_j = 0 \,|\, \eta))\right) \\
&= \mathrm{E}\left(\prod_{j=2}^{n} (1 - \Pr(e_j = 1 \,|\, \eta))\right). \quad (3)
\end{aligned}$$

Notice that $\Pr(e_j = 1 \,|\, \eta)$ is the probability that an edge exists between X_1 and X_j, conditioned on η. Consequently, $\xi = \Pr(e_j = 1 \,|\, \eta)$ is a random variable, depending on η. At the same time, it does not depend on j, as by de Finetti's theorem, the e_j are conditionally i.i.d, given η, so it does not matter which j is taken in $\xi = \Pr(e_j = 1 \,|\, \eta)$. Thus, we can continue (3) as

$$\Pr(X_1 \in \mathcal{I}_n) = \mathrm{E}\left(\prod_{j=2}^{n} (1 - \xi)\right) = \mathrm{E}\left((1 - \xi)^{n-1}\right). \quad (4)$$

We can now observe that $\xi \in [0, 1]$ and the function $g(x) = (1 - x)^n$ is convex in $[0, 1]$, so we may apply Jensen's inequality. Jensen's well known inequality says that for any random variable ζ and for any convex function g the inequality $\mathrm{E}\big(g(\zeta)\big) \geq g\big(\mathrm{E}(\zeta)\big)$ holds, which is a consequence of the definition of convexity. Thus, we can further continue (4), obtaining

$$\Pr(X_1 \in \mathcal{I}_n) = \mathrm{E}\left((1 - \xi)^{n-1}\right) \geq (1 - \mathrm{E}(\xi))^{n-1}.$$

Note that $\mathrm{E}(\xi) = \mathrm{E}(\Pr(e_j = 1 \,|\, \eta)) = \Pr(e_j = 1)$ is the probability that an edge exists between X_1 and X_j. By name invariance, this is the same probability for any two nodes, let p_n denote this common value. Thus,

$$\Pr(X_1 \in \mathcal{I}_n) \geq (1 - p_n)^{n-1}$$

follows. We know that there are $n - 1$ potential edges adjacent to each node, each with probabilty p_n. Therefore, despite the possible dependence of edges, the linearity of expectation implies the expected degree of each node under our conditions is $(n - 1)p_n$, which is also equal to $\overline{d}(n)$. We can then substitute $p_n = \overline{d}(n)/(n - 1)$, which yields

$$\Pr(X_1 \in \mathcal{I}_n) \geq \left(1 - \frac{\overline{d}(n)}{n - 1}\right)^{n-1},$$

implying

$$\mathrm{E}(I_n) = n \Pr(X_1 \in \mathcal{I}_n) \geq n \left(1 - \frac{\overline{d}(n)}{n - 1}\right)^{n-1}.$$

Assume now $\mathcal{M} \in \mathbf{BD}$, which means there is a constant C with $\overline{d}(n) \leq C$ for every n. Then

$$\left(1 - \frac{\overline{d}(n)}{n - 1}\right)^{n-1} \geq \left(1 - \frac{C}{n - 1}\right)^{n-1} \to \mathrm{e}^{-C},$$

so there exist constants $a > 0$ and $n_0 \in \mathbf{N}$, such that $\mathrm{E}(I_n) \geq an$ holds for every $n \geq n_0$.

Now take a sequence $\beta_n \in [0, 1]$ with $\beta_n \to 1$. We are going to show that the probability $\Pr(G_n$ is β_n-connected) cannot tend to 1, meaning $\mathcal{M} \notin \beta_n$–**CONN**.

Set $s_n = \Pr(I_n \leq (1 - \beta_n)n)$. Then $\Pr(G_n$ is β_n-connected) $\leq s_n$ must hold, since β_n-connectivity implies that there may be at most $(1 - \beta_n)n$ isolated nodes. Consider now the random variable $\gamma_n = n - I_n$. The definition of γ_n implies $\gamma_n \geq 0$ and $\mathrm{E}(\gamma_n) = n - \mathrm{E}(I_n)$. Therefore, $\mathrm{E}(\gamma_n) \leq (1 - a)n$ holds for $n \geq n_0$. Moreover, the definition also directly implies that the events $\{I_n \leq (1 - \beta_n)n\}$ and $\{\gamma_n \geq \beta_n n\}$ are equivalent. Thus, we can write, using Markov's inequality for nonnegative random variables:

$$s_n = \Pr(I_n \leq (1 - \beta_n)n) = \Pr(\gamma_n \geq \beta_n n) \leq$$

$$\leq \frac{\mathrm{E}(\gamma_n)}{\beta_n n} \leq \frac{(1 - a)n}{\beta_n n} = \frac{1 - a}{\beta_n}.$$

Since we know that $a > 0$ is a constant and $\beta_n \to 1$, therefore, there must exist a constant $b < 1$, such that $s_n \leq b$ holds for all large enough n. This, together with $\Pr(G_n$ is β_n-connected) $\leq s_n$, proves that the assumptions we made, that is, $\mathcal{M} \in \mathbf{GEOM}(loc, inv)$ and $\mathcal{M} \in \beta_n$–**CONN**, together imply $\mathcal{M} \notin \mathbf{BD}$, proving the theorem.

♠

As a corollary, we obtain that $\mathbf{GEOM}(loc, inv)$ is smaller than $\mathbf{GEOM}(loc)$ and $\mathbf{GEOM}(inv)$.

Corollary 2 $\mathbf{GEOM}(loc, inv) \not\subseteq \mathbf{GEOM}(loc)$ *and* $\mathbf{GEOM}(loc, inv) \not\subseteq \mathbf{GEOM}(inv)$.

Proof. Let $\mathcal{M} = (G_n;\ n \in \mathbf{N})$ be a model in which G_n is chosen unformly at random from the set of all connected graphs with maximum degree at most 3. It follows from this construction that $\mathcal{M} \in \mathbf{BD} \cap \mathbf{CONN}$, implying $\mathcal{M} \in \mathbf{BD} \cap \beta_n$–**CONN** for any β_n. Then Theorem 3 implies $\mathcal{M} \notin \mathbf{GEOM}(loc, inv)$. Since, naturally, $\mathcal{M} \in \mathbf{GEN}$, therefore, it follows that $\mathbf{GEOM}(loc, inv) \not\subseteq \mathbf{GEN}$. As we know from

Corollary 1 that $\mathbf{GEOM}(loc) \simeq \mathbf{GEOM}(inv) \simeq \mathbf{GEN}$, we obtain $\mathbf{GEOM}(loc, inv) \not\simeq \mathbf{GEOM}(loc)$ and $\mathbf{GEOM}(loc, inv) \not\simeq \mathbf{GEOM}(inv)$.

♠

4 An Application

In this application example we model a mobile wireless ad hoc network, that is, a network in which wireless nodes communicate to each other directly, without a supporting infrastructure, The initial position of each node is chosen in the following way. Let P be a probability measure over a planar domain D. First we choose k pivot points independently at random, using P. Then the actual node positions are generated such that each potential node is chosen independently at random from P, but it is kept only if it is within a given distance d_0 to at least one of the random pivot points, otherwise it is discarded. Note that this way of generating the nodes makes them dependent, as the non-discarded ones cluster around the random pivot points, thus modeling a clustered, non-independent node distribution.

The mobility of the nodes in this example is modeled in the following way. Over some time horizon T_n, that may depend on n, the number of nodes, each node moves along a random curve from its initial position with a constant speed v_0. The curve is chosen from a set $\mathcal{C}$ of available potential trajectories in D. For simplicity, it is assumed that each curve can be identified by a real parameter. This parameter is chosen using a probability distribution $Q_{x,y}$ that depends on the initial position (x, y) of the node. Then the randomly obtained curve is shifted so that its startpoint coincides with the random initial position of the node and then the node will move along this random trajectory.

Let $d(x, y)$ be a nonnegative real valued function over $D \times D$, with the only restriction that $d(x, x) = 0$ holds for any x. This function is intended to measure "radio distance" in D. The assumption is that whenever $d(x, y)$ is small enough, then two nodes positioned at x and y can receive each others' transmissions. The function $d(x, y)$, however, does not have to satisfy the usual distance axioms, it may reflect complex radio propagation characteristics, such as expected attenuation and fading, it may account for the heterogeneity of the terrain, for propagation obstacles etc. We may also include random effects, making $d(x, y)$ a random variable, reflecting special conditions of interest, such as the random presence of eavesdroppers that can trigger the inhibition of certain links. We assume, however, that if there is randomness in $d(x, y)$, then it is independent of the other random variables in the model.

Let t_n and r_n be further parameters that may also depend on the number n of nodes. We now define the links of the network, as follows. Consider two nodes with initial position vectors $X_1(0), X_2(0)$, respectively. As they move along their random trajectories, their positions at time t is denoted by $X_1(t), X_2(t)$, respectively. The two nodes are considered connected by a link, if there is a closed subinterval of length at least t_n within the time horizon $[0, T_n]$, such that $d(X_1(t), X_2(t)) \leq r_n$ holds for every time t within the subinterval[8], with the possibly complicated radio distance.

Now the question is this: for given P, D, $\mathcal{C}$, $Q_{x,y}$ and $d(x, y)$, and for the described way of dependent node genaration, can we somehow choose the model parameters k, d_0, v_0, T_n, t_n and r_n, such that the arising random graph is a.a.s. connected, while the expected average degree in the graph remains bounded?

We believe that it would be rather hard to answer such a question with a direct analysis for arbitrary complex choices of P, D, $\mathcal{C}$ $Q_{x,y}$ and $d(x, y)$. On the other hand, with our general results it becomes quite straightforward, showing the strength of the results.

Let us choose the model domain S as a 3-dimensional phase space, in which each node is represented by a point such that the first two coordinates describe the intial position of the node and the last coordinate encodes which random trajectory was chosen from $\mathcal{C}$ for the node. Let $X_1, X_2, \ldots$ be the representations of the nodes in this phase space.

We can now check that, for any n, the joint distribution of $X_1, \ldots, X_n$ is invariant to re-indexing them. The reason is that both the initial positions and the trajectory choices are generated by processes in which the indices do not play any role. Therefore, the model is *name invariant*. Interestingly, this remains true despite having a lot of dependencies among the nodes: the initial positions of different nodes are not independent (due to clustering), and the trajectory of a given node is also not independent of its initial position, as it is drawn from a probability distribution that may depend on the location. Through this, the trajectories and initial positions of different nodes also become dependent, making their whole movement dependent. Yet, the model is still name invariant.

Let us now consider the links. As defined above, two nodes are considered connected if during their movement over the time horizon $[0, T_n]$ there is a subinterval of time, of length at least t_n, such that they remain within "radio distance" $\leq r_n$ during the entire subinterval. The radio distance, however, may be very different from the Euclidean distance, it may be described by an arbitrary function that may account for complex propagation characteristics, attenuation, obstacles, and it may also contain independent randomness.

Given some possibly complicated radio distance $d(x, y)$ and the node generation and movement process with possibly complex trajectories, it may not be easy to compute whether a link actually exists between two nodes according to the above definition. On the other hand, for us it is enough to note that once the phase space representations X_i, X_j of any two nodes are given, plus the realization of the independent randomness of the distance, they together determine whether a link exists between the two nodes or not. The reason is that the initial positions and the trajectories, given in the phase space representation, fully determine the movement of the nodes. Once this is known, it determines, along with the realization of the independent randomness of the distance function, whether the link definition is satisfied, i.e., if there is a subinterval of length $\geq t_n$ in $[0, T_n]$, such that the nodes stay within radio distance $\leq r_n$ during the entire subinterval. To actually compute it may not be easy for a sophisticated case, but for our purposes it enough to know that it is *determined* by the listed factors, without knowing anything about the other nodes. This implies that the model is *local*.

[8]The motivation is that the nodes should be within range at least for the time of sending a packet.

Thus, we have established that, for any choice of the parameters, the problem can be described by a model that is in $\mathbf{GEOM}(loc, inv)$, Then this model cannot be in $\mathbf{BD} \cap \mathbf{CONN}$, since we know from Theorem 3 that $\mathbf{BD} \cap \beta_n\mathbf{-CONN} \cap \mathbf{GEOM}(loc, inv) = \emptyset$ holds for any choice of $\beta_n \to 1$, including $\beta_n \equiv 1$. Thus, in our example it is impossible to keep the expected average degree bounded and achieveing a.a.s. connectivity at the same time. With this we could cut through a lot of complexity that would otherwise arise with the direct analysis of the specific model.

5 Conclusion and Open Problems

Our research has been motivated by the fact that many different random graph constructions are used to model large real life graphs, but often it is not clear how the strength of the different models compare to each other, e.g., when does it hold that a certain model property implies another. We have set up a general framework to compare the strength of various random graph model classes, and presented some results about the equality, inequality and proper containment of these classes.

Since we have just initiated this line of investigation, many questions remain open. Let us mention two examples that seem interesting and nontrivial.

Open problem 1. One can easily see from the definition that $\mathbf{GEOM}(loc, free) \preceq \mathbf{IDE}$. That is, in local geometric models with independent node variables the disjoint edges are independent. Is the converse true, i.e., can we represent any $\mathcal{M} \in \mathbf{IDE}$ by a local geometric model with independent node variables?

Open problem 2. Is it true that in every local and name invariant geometric model the edges are positively correlated? In other words, does $\mathbf{GEOM}(loc, inv) \preceq \mathbf{POS}$ hold? Or does at least $\mathbf{GEOM}(loc, free) \preceq \mathbf{POS}$ hold? If it were true, this would have important consequences for geometric random graph models.

References

Bornholdt, S. & Shuster, H.G. (2003), *Handbook of Graphs and Networks — From the Genome to the Internet*, Wiley-VCH.

de Finetti, B. (1931), 'Funzione Caratteristica di un Fenomeno Aleatorio', *Atti della R. Academia Nazionale dei Lincei, Serie 6, Classe di Scienze Fisiche, Mathematice e Naturale,* **4**, 251–299.

Faragó, A. (2007), 'On the Fundamental Limits of Topology Control in Ad Hoc Networks', *Algorithmica*, **49**, 337–356.

Franceschetti, M. & Meester, R. (2007), *Random Networks for Communication*, Cambridge University Press.

Kallenberg, O. (2005), *Probabilistic Symmetries and Invariance Principles*, Springer.

Penrose, M. (2003), *Random Geometric Graphs*, Oxford University Press.

Stoyanov, J.M. (1987), *Counterexamples in Probability*, Wiley.