International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2018, Submitted in January 2018, iajesm2014@gmail.com

Random Graphs: Theory and Applications

B.V. Manjunatha, Assistant Professor, Department of Mathematics, Government first grade women's college Tumkur, Karnataka (India), E-mail address: bvm.maths@gmail.com

Abstract

The theory of random graphs deals with asymptotic properties of graphs equipped with a certain probability distribution; for example, it studies how the component structure of a uniform random graph evolves as the number of edges increases. Since the foundation of the theory of random graphs by Erdos and R ″ enyi five ′ decades ago, various random graph models have been introduced and studied. Graph theory has meanwhile found its way into other sciences as a rich source of models describing fundamental aspects of a broad range of complex phenomena. This article is a gentle introduction to the theory of random graphs and its recent developments (with focus on the phase transition and critical phenomena, a favourite topic of the first author) and applications. This is an extended version of the article entitled "Random Graphs: from Nature to Society" published in Seoul Intelligencer, a special issue of the Mathematical Intelligencer, on the occasion of International Congress of Mathematicians in Seoul in 2014.

Keywords: Random graph, theory, application

Introduction:

Random graph inference is an active, interdisciplinary area of current research, bridging combinatorics, probability, statistical theory, and machine learning, as well as a wide spectrum of application domains from neuroscience to sociology. Statistical inference on random graphs and networks, in particular, has witnessed extraordinary growth over the last decade: see, for example, Goldenberg et al. (2010) and Kolaczyk (2009) for a discussion of the considerable applications in recent network science of several canonical random graph models.

Of course, combinatorial graph theory itself is centuries old—indeed, in his resolution to the problem of the bridges of K¨onigsberg, Leonard Euler first formalized graphs as mathematical objects consisting of vertices and edges. The notion of a random graph, however, and the modern theory of inference on such graphs, is comparatively new, and owes much to the pioneering work of Erd″os, R´enyi, and others in the late 1950s. E.N. Gilbert's short 1959 paper (Gilbert, 1959) considered a random graph for which the existence of edges between vertices are independent Bernoulli random variables with common probability p; roughly concurrently, Erd″os and R´enyi provided the first detailed analysis of the probabilities of the emergence of certain types of subgraphs within such graphs (Erd″os and R´enyi, 1960), and today, graphs in which the edges arise independently and with common probability p are known as Erd″os-R´enyi (or ER) graphs

The Erd″os-R´enyi (ER) model is one of the simplest generative models for random graphs, but this simplicity belies astonishingly rich behavior (see Alon and Spencer, 2008; Bollob´as et al., 2007). Nevertheless, in many applications, the requirement of a common connection probability is too stringent: graph vertices often represent heterogeneous entities, such as different people in a social network or cities in a transportation graph, and the connection probability pij between vertex i and j may well change with i and j or depend on underlying attributes of the vertices. Moreover, these heterogeneous vertex attributes may not be observable; for example, given the adjacency matrix of a Facebook community, the specific interests of the individuals may remain hidden. To more effectively model such real-world networks, we consider latent position random graphs (Hoff et al., 2002). In a latent position graph, to each vertex i in the graph there is associated an element xi of the so-called latent space X , and the probability of connection pij between any two edges i and j is given by a link or kernel function $\kappa : X \times X \to [0, 1]$. That is, the edges are generated independently (so the graph is an independent-edge graph) and $pij = \kappa(xi , xj)$.

In any latent position graph, the latent positions associated to graph vertices can themselves be random; for instance, the latent positions may be independent, identically distributed random variables with some distribution F on R d . The well-known stochastic blockmodel (SBM), in which each vertex belongs to one of K subsets known as blocks, with connection

probabilities determined solely by block membership (Holland et al., 1983), can be represented as a random dot product graph in which all the vertices in a given block have the same latent positions (or, in the case of random latent positions, an RDPG for which the distribution F is supported on a finite set). Despite their structural simplicity, stochastic block models are the building blocks for all independent-edge random graphs; in Wolfe and Olhede (2013), the authors demonstrate that any independent-edge random graph can be well-approximated by a stochastic block model with a sufficiently large number of blocks. Since stochastic block models can themselves be viewed as random dot product graphs, we see that suitably high-dimensional random dot product graphs can provide accurate approximations of latent position graphs (Tang et al., 2013), and, in turn, independent-edge graphs. Thus, the architectural simplicity of the random dot product graph makes it particularly amenable to analysis, and its near-universality in graph approximation renders it expansively applicable. In addition, the cornerstone of our analysis of random dot product graphs is a set of classical probabilistic and linear algebraic techniques that are useful in much broader settings, such as random matrix theory. As such, the random dot product graph is both a rich and interesting object of study in its own right and a natural point of departure for wider graph inference.

The ambition and scope of our approach to graph inference means that mere upper bounds on discrepancies between parameters and their estimates will not suffice. Such bounds are legion. In our proofs of consistency, we improve several bounds of this type, and in some cases improve them so drastically that concentration inequalities and asymptotic limit distributions emerge in their wake. We stress that aside from specific cases (see F¨uredi and Koml´os, 1981; Tao and Vu, 2012; Lei, 2016), limiting distributions for eigenvalues and eigenvectors of random graphs are notably elusive. For the adjacency and Laplacian spectral embedding, we discuss not only consistency, but also asymptotic normality, robustness, and the use of the adjacency spectral embedding in the nascent field of multi-graph hypothesis testing. We illustrate how our techniques can be meaningfully applied to thorny and very sizable real data, improving on previously state-of-the-art methods for inference tasks such as community detection and classification in networks. What is more, as we now show, spectral graph embeddings are relevant to many complex and seemingly disparate aspects of graph inference.

Review of Literature:

A bird's-eye view of our methodology might well start with the stochastic blockmodel. For an SBM with a finite number of blocks of stochastically equivalent vertices, in Sussman et al. (2012) and Fishkind et al. (2013), we establish that k-means clustering of the rows of the adjacency spectral embedding accurately partitions the vertices into the correct blocks, even when the embedding dimension is misspecified or the number of blocks is unknown.

Furthermore, in Lyzinski et al. (2014) and Lyzinski et al. (2017) we give a significant improvement in the misclassification rate, by exhibiting an almost-surely perfect clustering in which, in the limit, no vertices whatsoever are misclassified. For random dot product graphs more generally, we show in Sussman et al. (2014) that the latent positions are consistently estimated by the embedding, which then allows for accurate learning in a supervised vertex classification framework. In Tang et al. (2013), these results are extended to more general latent position models, establishing a powerful universal consistency result for vertex classification in general latent position graphs, and also exhibiting an efficient embedding of vertices which were not observed in the original graph. In Athreya et al. (2016) and Tang and Priebe (2016), we supply distributional results, akin to a central limit theorem, for both the adjacency and Laplacian spectral embedding, respectively; the former leads to a nontrivially superior algorithm for the estimation of block memberships in a stochastic block model (Suwan et al., 2016), and the latter resolves, through an elegant comparison of Chernoff information, a long-standing open question of the relative merits of the adjacency and Laplacian graph representations.

Morever, graph embedding plays a central role in the foundational work on hypothesis testing of Tang et al. (2017a) and Tang et al. (2017b) for two-sample graph comparison: these papers provide theoretically justified, valid and consistent hypothesis tests for the semiparamatric

problem of determining whether two random dot product graphs have the same latent positions and the nonparametric problem of determining whether two random dot product graphs have the same underlying distributions. This, then, yields a systematic framework for determining statistical similarity across graphs, which in turn underpins yet another provably consistent algorithm for the decomposition of random graphs with a hierarchical structure Lyzinski et al. (2017). In Levin et al. (2017), distributional results are given for an omnibus embedding of multiple random dot product graphs on the same vertex set, and this embedding performs well both for latent position estimation and for multi-sample graph testing. For the critical inference task of vertex nomination, in which the inference goal is to produce an ordering of vertices of interest (see, for instance Coppersmith, 2014), we find in Fishkind et al. (2015a) an array of principled vertex nomination algorithms —-the canonical, maximum likelihood and spectral vertex nomination schemes—and a demonstration of the algorithms' effectiveness on both synthetic and real data.

In Lyzinski et al. (2016b) the consistency of the maximum likelihood vertex nomination scheme is established, a scalable restricted version of the algorithm is introduced, and the algorithms are adapted to incorporate general vertex features. Overall, we stress that these principled techniques for random dot product graphs exploit the Euclidean nature of graph embeddings but are general enough to yield meaningful results for a wide variety of random graphs. Because our focus is, in part, on spectral methods, and because the adjacency matrix A of an independent-edge graph can be regarded as a noisy version of the matrix of probabilities P (Oliveira, 2009), we rely on several classical results on matrix perturbations, most prominently the Davis-Kahan Theorem (see Bhatia (1997) for the theorem itself, Rohe et al. (2011) for an illustration of its role in graph inference, and Yu et al. (2015) for a very useful variant). We also depend on the aforementioned spectral bounds in Oliveira (2009) and a more recent sharpening due to Lu and Peng (Lu and Peng, 2013). We leverage probabilistic concentration inequalities, such as those of Hoeffding and Bernstein (Tropp, 2015). Finally, several of our results do require suitable eigengaps for P and lower bounds on graph density, as measured by the maximum degree and the size of the smallest eigenvalue of P. It is important to point out that in our analysis, we assume that the embedding dimension d of our graphs is known and fixed. In real data applications, such an embedding dimension is not known, and in Section 6.3, we discuss approaches (see Chatterjee, 2015; Zhu and Ghodsi, 2006) to estimating the embedding dimension. Robustness of our procedures to errors in embedding dimension is a problem of current investigation.

Random Graph Models

A random graph is obtained by starting with a set of n vertices and adding edges between them at random. Different random graph models produce different probability distributions on graphs. The most commonly studied model, usually called the Erdos-Renyi graphs, is written as $G_{n,p}$, where n is the num- ber of nodes in the graph and p is the probability of any edge existing between any pair of nodes. This probability for one edge is independent of the existence of any other edge in the graph. Based on these assumptions we can find out that the average degree of $G_{n,p}$ is $z = \frac{p}{n-1} \approx \frac{p}{n}$ as n is large (n>>1). Also, the probability of a node having degree k is given by $p_k = ({}^{N}C_k)p^k(1-p)^{N-k} = \frac{e^{-z}z^k}{k!}$.

A closely related model, $G_{n,m}$ defines the set of graphs having n vertices and m randomly selected edges. Still another model of random graphs is a random graph with a given arbitrary probability distribution of the degrees of their vertices. In all respects other than their degree distribution, these graphs are assumed to be entirely random. This means that the degrees of all vertices are independent identically distributed random integers drawn from a specified distribution. For a given choice of these degrees, also called the "degree sequence", the set of random graphs having the degree sequence is called a Microcanonical Ensemble.

Microcanonical Ensemble

In studying the properties of random graphs, graph theorists often concentrate on the limit behavior of random graphsthe values that various probabilities converge to as n grows very large. In such cases, a Microcanonical Ensemble is a set of all large graphs having the same

degree sequence that matches as closely as possible to the desired degree probability distribution. Properties of such graphs are calculated by averaging over the whole ensemble of graphs of the given degree sequence.

Phase Transition

One of the most interesting aspects of this addition of the edges to form the random graph is the Phase Transition. There are two distinct phases in the formation of random graphs. Initially, the graph is disconnected and later, after addition of a certain number of edges, the graph becomes largely connected. Largely connected need not mean fully connected, it only means a large majority of the nodes is connected. Here comes the concept of Giant Components. Giant components are large connected components of a random graph, whose size is proportional to the size of the whole graph, i.e. O(n). So it increases linearly as the size of the graph increases. The emergence of GC in a evolving random graph marks the transition of the graph to the connected phase. Erdos and Renyi found out that there is a sharp threshold for the emergence of giant components, which is as follow: [7]

- If $p = c/n$ and $c < 1$ then, when n is large, most of the connected components of the graph are small, with the largest having only $O(\log n)$ vertices.
- In contrast if $c > 1$ there is a constant $\Theta(c) > 0$ so that for large n the largest component has $\sim \Theta(c)n$ vertices and the second largest component is $O(\log n)$.

Giant Components

Giant Components is perhaps the most studied phenomenon in the field of random graphs is the be- havior of the size of the largest component in $G_{n,p}$. The major question on which we will be concentration in this discussion is that whether there can exist multiple giant components in a large random graph or not. For that purpose let us first understand the definitions of the terms to be used, then we prove that in the thermodynamic limit multiple giant components cannot exist.

Multiple Giant Components

One of the major question that arises in relation to giant components is that whether there can exist multiple giant components in a large random graph or not. So let us try to find out whether two giant components can exist is a random graph. That is given a ER random graph $G_{n,p}$ of n nodes, what is the probability that there exist two giant components GC1 (size n1) and GC2 (size N2). We are using the $G_{n,p}$ model, so we want to find out that what is

connected by the edges that are randomly *thrown* on the graph.

$$P(GC1 \text{ and } GC2 \text{ not connected by the edge}) = 1 - P(GC1 \text{ and } GC2 \text{ gets connected by the edge})$$

$$= 1 - \frac{N1 * N1}{n(n-1)} \quad [1]$$

$$Total\ number\ edges\ in\ G_{n,p} = {}^{n}C_2.p \quad [2]$$

Therefore,

$$P(none\ of\ those\ edges\ connect\ GC1\ and\ GC2) = \left(1 - \frac{N1 * N1}{n(n-1)}\right)^{{}^{n}C_2.p} \quad [3]$$

Here we have taken the following assumptions.

- N1 = O(n) and N2 = O(n) but n >> N1,N2 >> m
- N1 and N2 are so big that addition of a node to n or addition of an edge from N1 to N2 does not make any difference in the probabilities.

Now let us try to analyze what happens to the probability at the thermodynamic limit, i.e. $n \to \infty$.

$$Let\ L = Lim_{n\to\infty} = \left[\left(1 - \frac{N1 * N1}{n(n-1)}\right)^{{}^{n}C_2.p}\right] \quad [4]$$

Now, at $n \to \infty$, ${}^{n}C_2 \approx \frac{n^2}{2}$. Therefore,

$$L = Lim_{n\to\infty}\left[\left(1 - \frac{N1.N1}{\frac{n^2}{2}}\right)^{\frac{n^2}{2}.p}\right] \quad [5]$$

$$L = \left[Lim_{n\to\infty}\left(1 - \frac{N1.N1}{\frac{n^2}{2}}\right)^{\frac{n^2}{2}}\right]^{p} \quad [6]$$

$$L = e^{-N1.N2.p} \quad [7]$$

For random graphs $G_{n,p}$ the average degree $z = (n-1).p$, i.e. $z \approx n.p$. Also we know $N1 = O(n)$, hence $N1 = \delta_1.n$. Similarly, $N1 = \delta_2.n$. Substituting these we get,

$$L = e^{-N1.N2.p} = e^{-\delta_1.\delta_1.n^2.\frac{c}{n}} \quad [8]$$

$$L = e^{-const.n} \quad [9]$$

At thermodynamic limit, $Lim_{n\to\infty} L = 0$. Therefore we can conclude that as the size of the ER random graph increase to infinity, the probability of having 2 giant components tends to zero.

the probability that these two components will not get

Existence of a node in GC

At the point of formation of the single giant component, the size of the component is $O(n^{\frac{2}{3}})$.
Let us try to analyze the probability of a node being in the giant component. Let u be the probability that the node is not in a giant component. Probability that all its k neighbors are not in giant component is u^k.
Now, if a node is in a giant component, then it implies all its neighbors are not in the giant component.
Prob of one node = Prob of having k neighbors X Prob of all k neighbors not in GC

$$u = \sum_{k=0} p_k . u^k \quad [10]$$

where p_k = probability of a node having k neighbors = $({}^{N}C_k)p^k(1-p)^{N-k} = \frac{e^{-z}.z^k}{k!}$

$$u = \sum_{k=0} \frac{e^{-z}.z^k}{k!}.u^k \quad [11]$$

$$u = e^{-z}.\sum_{k=0} \frac{(z.u)^k}{k!} = e^{-z}.e^{z.u} \quad [12]$$

$$u = e^{-z(1-u)} \quad [13]$$

Therefore s = probability of node not being in GC = $1 - u$

$$s = 1 - e^{-z.s} \quad [14]$$

The first non-zero solution of this equation is the required probability.

Existence of GC in a generalized random graph of given degree sequence

In the paper The size of the giant component of a random graph with a given degree sequence by Molloy & Reed, they have suggested the following:

Given a sequence of nonnegative real numbers $\lambda_0, \lambda_1, \lambda_2, \ldots$ which sum to 1, a random graph having approximately $\lambda_i n$ vertices of degree i will have a giant component at the thermodynamic limit if $\sum i(i-2)\lambda_i > 0$.

This essentially means that given a degree sequence $k_0, k_1, k_2, \ldots$, a large random graph $(n \to \infty)$having that degree sequence will have a giant component if $\sum k_i(k_i - 2) > 0$.

This can be understood in an intuitive manner. If we are trying to traverse a the network like a graph by maintaining a list of unexplored nodes, then for a giant component to exist we must ensure that the list does not become empty i.e. the connected component can go on expanding. Now when we come to a node i having degree k_i, we now have k_i new nodes to traverse, which we have got at the cost of traversing

the node i. So in the list of unexplored nodes increases by $(k_i - 1) - 1 = k_i - 2$.
Now we can argue that probability of reaching a node of degree k_i is k_i times the probability of reaching a node of degree 1 (because it can be reached by k different edges). Therefore,

$$P(\text{reaching node of deg } k_i) = k_i P(\text{reaching node of deg } 1) = k_i.const$$

Hence we can say that $\sum p_k(k-2) > 0$ ensures that the list of unexplored nodes will never be empty.
Let the sum be S.

$$S = \sum p_k(k-2) \approx \sum (k_i.const)(k_i - 2) = const. \sum k_i(k_i - 2) > 0 \quad [15]$$

$$\sum k_i(k_i - 2) > 0 \quad [16]$$

This is the result we have.

Conclusion

From the above discussions we can conclude that in the asymptotic case, ER graph of the form $G_{n,p}$ cannot have more than one giant components. Also the probability of a node

being in the giant compo- nent is $s = 1 - e^{-z.s}$. And given a degree sequence $k_0, k_1, k_2, \ldots$, a large random graph at thermodynamic limit having that degree sequence will have a giant component if $\sum k_i(k_i - 2) > 0$.

References

1. "On Random Graphs" - Erdos P., and A. Renyi, Publicationes Mathematicae 6, pp.290-297 (1959)
2. "On the Evolution of Random Graphs" - Erdos P., and A. Renyi, Hungarian Academy of Sciences 5, pp. 17-61 (1960)
3. "Random graphs with arbitrary degree distribution and their applications" - M. E. J. Newman, S. H. Strogatz and D. J. Watts, Reviews of Modern Physics, 2002
4. "A critical point for random graphs with a given degree sequence" - M. Molloy & B. Reed, Random Structures and Algorithms 6, 161-179 (1995)
5. "The size of the giant component of a random graph with a given degree sequence" - M. Molloy & B. Reed, Combinatorics, Probability and Computing 7, 295-305(1998).
6. Random Graphs on Wikipedia - http : //en.wikipedia.org/wiki/Random$_g$raph
7. Random Graph Dynamics - Book by Rick Durrett, Cornell U., Published by Cambridge U. Press, October 2006 http://www.math.cornell.edu/ durrett/RGD/fch1.pdf