Random intersection graphs when $m = \omega(n)$: an equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models

James Allen Fill *
Edward R. Scheinerman †
Karen B. Singer-Cohen ‡

August, 1998

ABSTRACT

When the random intersection graph $G(n, m, p)$ proposed by Karoński, Scheinerman, and Singer-Cohen in [8] is compared with the independent-edge $G(n, p)$, the evolutions are different under some values of m and equivalent under others. In particular, when $m = n^\alpha$ and $\alpha > 6$, the total variation distance between the graph random variables has limit 0.

Key Words: Random graphs, intersection graphs, total variation distance, threshold

AMS Subject Classifications: 05C80 Random graphs

*Department of Mathematical Sciences, The Johns Hopkins University, email: jimfill@jhu.edu

†Department of Mathematical Sciences, The Johns Hopkins University, email: ers@jhu.edu

‡Department of Mathematics, Wellesley College, email: kcohen@wellesley.edu

1 Introduction

In 1945 E. Marczewski proved that every graph can be represented by a list of sets, where one set is associated with each vertex([10]). In this representation, pairs of sets with nonempty intersections correspond to edges in the graph. As part of the study of such *intersection graphs*, it is natural to ask what sorts of graphs would be most likely to arise if the list of sets is generated randomly.

One such model, $G(n, m, p)$, in which the elements of the sets are chosen independently from a universal set, is introduced in [8]. As random graph theory concerns largely the "evolutionary" features of a model when its parameters values are changed, this paper examines the evolution of $G(n, m, p)$. In particular, it compares the evolution with that of the independent-edge random graph model $G(n, p)$, using total variation distance as a measure of similarity between the graph-valued random variables.

It should be noted that the primary conceptual and practical difference between the $G(n, m, p)$ random intersection graph model and its well-studied random graph predecessors is in the "location" of the randomness. While earlier models focus on assigning edges to the graph randomly, this new model assigns a random structure (a finite set) to each vertex of the graph. The action thus centers on the vertices, with the edge set determined by vertex outcomes. This allows for modeling of more complex situations, in which relationships between objects (edges between vertices) are dependent on properties of those particular objects (vertices). As this paper will show, the strength of this dependence can be controlled by adjusting one of the

parameters of $G(n, m, p)$.

Some of the asymptotic and probabilistic notation of this paper is introduced below.

a.s.	"almost surely" in the graph theoretic sense, i.e., with probability tending to 1 as $n \to \infty$
$\mathbf{E}X$	expectation of the random variable X
$f(n) = o(g(n))$	$f(n)/g(n) \to 0$ as $n \to \infty$
$f(n) \ll g(n)$	$f(n)/g(n) \to 0$ as $n \to \infty$ (equivalently, $f(n) = o(g(n))$)
$f(n) = \omega(g(n))$	$f(n)/g(n) \to \infty$ as $n \to \infty$
$f(n) \gg g(n)$	$f(n)/g(n) \to \infty$ as $n \to \infty$ (equivalently, $f(n) = \omega(g(n))$)
$f(n) = O(g(n))$	$f(n)/g(n) \leq c$ as $n \to \infty$, where $0 < c < \infty$ is constant
$f(n) \asymp g(n)$	$c_1 \leq f(n)/g(n) \leq c_2$ as $n \to \infty$, where $0 < c_1 < c_2 < \infty$ are constants
$f(n) \sim g(n)$	$f(n)/g(n) \to 1$ as $n \to \infty$ (equivalently, $f(n) = g(n)(1 + o(1))$)

The following well-known approximations and bounds are used in this paper.

1. If $mp \to 0$, then

$$\begin{aligned} (1-p)^m &\geq 1 - mp, \text{ and} \\ (1-p)^m &\leq 1 - mp + \binom{m}{2}p^2, \text{ implying that} \\ (1-p)^m &= 1 - mp + O(m^2p^2) = 1 - (1 - o(1))mp. \end{aligned}$$

This last estimate, in turn, implies that $1 - (1-p)^m \sim mp$. The

bounds can be proved using the binomial expansion of $(1-p)^m$ and the fact that the resulting alternating terms are decreasing in absolute value.

2. If $0 < a < 1$, $0 < b < 1$, and $c > 1$, then $(1-a)^{bc} > (1-ab)^c$.

3. For all $m, p > 0$, $e^{-mp} - (1-p)^m \leq e^{-mp}\frac{mp^2}{2(1-p)^2}$.

4. If $mp^2 \to 0$, then $(1-p)^m \sim e^{-mp}$. This is shown by using the expansion for $\ln(1-p)$ when $|p| < 1$ and the fact that $1 - e^{-a} \leq a$ for $a \geq 0$.

2 The Models

2.1 $G(n,p)$

Most random graph work has been done on the $G(n,p)$ model, one of the original models established and developed by P. Erdős and A. Rényi starting in the late 1950s ([4],[5]). In this random graph model, the parameter n refers to the number of vertices on which the graph is based, and $p = p(n)$ gives the probability that there will be an edge between any two particular vertices. The event of edge appearance is independent for each of the possible $\binom{n}{2}$ places for edges. Thus the probability of obtaining a particular graph instance $G_{n,p}$ is determined completely by the numbers of edges and of non-edges for the configuration. This independence structure facilitates probability computations for certain kinds of events.

2.2 $G(n, M)$

Another model proposed by Erdős and Rényi is the fixed-size model $G(n, M)$. In this model, a graph is chosen uniformly at random from the set of all graphs on n vertices that have exactly M edges.

2.3 The New Model: $G(n, m, p)$

2.3.1 Parameters and Definitions

To generate a graph $G_{n,m,p}$ in the $G(n, m, p)$ model, let n and m be positive integers, and let $p \in [0, 1]$. Each of n vertices is assigned a random subset from a fixed set of m elements, with an edge arising between two vertices when their sets have at least one common element. The random subsets are picked by allowing each vertex to flip a p-weighted coin for each possible of the m elements, independently, to decide whether or not to include it in the set. The probability that a fixed vertex chooses a particular set S is thus $p^s(1-p)^{m-s}$, where $s = |S|$. These events are independent for different vertices. Formally,

1. For each $1 \leq i \leq n$, let S_i be a randomly generated subset of $[m] := \{1, 2, 3, \ldots, m\}$, where each element of $[m]$ is included in S_i independently with probability p. The random subsets $S_1, S_2, \ldots, S_n$ are assumed to be mutually independent.

2. Let $G = G_{n,m,p}$ be the random intersection graph with $V(G) = [n] = \{1, 2, 3, \ldots, n\}$ and $E(G) = \{(i, j) : i, j \in V, S_i \cap S_j \neq \emptyset\}$.

For the current work, m is restricted to be a function of n that is of the form

$$m = \lfloor n^\alpha \rfloor \text{ for some fixed, real } \alpha > 0.$$

This will be written somewhat casually in the sequel as $m = n^\alpha$. When m is taken to be much larger than this, the thresholds are as those for the Erdős–Rényi model, and as such are unlikely to yield new insights. The designated function $m = n^\alpha$ offers a controllable amount of deviation from the standard models, while allowing for a natural progression from sparse to dense graphs as p is increased.

2.3.2 A Matrix Description of the Model: $R(n, m, p)$

Each list of sets describing a random intersection graph instance $G_{n,m,p}$ may alternatively be represented by an $n \times m$ 0-1 matrix, referred to as $R_{n,m,p}$, or the *representation matrix*, an element in the probability space $R(n, m, p)$. Each row of the representation matrix corresponds to a vertex of the graph, and each column to an element of the universal m-set. The matrix entries are defined as follows:

$$R_{ij} = \begin{cases} 1 & \text{if } j \in S_i \\ 0 & \text{if } j \notin S_i. \end{cases}$$

Each entry of the matrix is thus independently 1 with probability p, and 0 otherwise. The relationship between $R_{n,m,p}$ and $G_{n,m,p}$ can be stated explicitly by specifying how one would find the graph $G = \Psi(R)$, given R. Indeed, $\Psi(R)$ is the graph on $V = [n]$ with

$$E = \{ij : i, j \in V, \text{ rows } i \text{ and } j \text{ of } R \text{ have a 1 in some common column}\}.$$

	elt_1	elt_2	elt_3	elt_4
v_1	1	0	0	0
v_2	1	0	1	0
v_3	1	0	1	1

	elt_1	elt_2	elt_3	elt_4
v_1	1	0	0	1
v_2	1	0	1	0
v_3	0	0	1	1

Figure 1: Two possible matrices leading to the triangle graph.

In other words there is an edge between i and j in G whenever the dot product of rows i and j of R is nonzero. In the matrix framework it is helpful to visualize the graph as arising through a set of cliques. (A clique in this paper is a set of pairwise adjacent vertices, not necessarily maximal.) A column with at least one 1 generates a clique in the graph, since all vertices that have 1's in a particular column are pairwise adjacent in the graph. Cliques generated by an individual column in this way will sometimes be referred to as *column cliques*. One may think of columns as picking vertices, independently, for their cliques. The union of all edges generated by such column cliques then forms the edge set of $G_{n,m,p}$. In general, a set of cliques that covers the edges of a graph G is called a *clique cover* for G.

More than one matrix can generate the same graph G. An example of two matrices that would generate a triangle on three vertices is given in Figure 1. The probability of obtaining a particular graph G on n vertices in a random trial is the sum of the probabilities of obtaining the various matrices that lead to that graph. For the probability spaces of $R(n, m, p)$

and $G(n, m, p)$, the function $\Psi : R(n, m, p) \to G(n, m, p)$ can be used to describe the relation between the respective probability measures. If $P_R(\cdot)$ is the probability measure in the space of matrices, then the measure in $G(n, m, p)$ is the indirect measure $P = P_R(\Psi^{-1})$. Note that for a particular graph $g \in G(n, m, p)$ to be generated, there are certain combinations of column types that must appear somewhere in the random matrix, and certain column types that are not allowed, as they would create extra, unwanted edges. Therefore, considering the m columns for the elements as multinomial trials, the probability that the column set fits the necessary constraints can be written as a sum of multinomial probability expressions. As a simple exercise, one may compute the probability of an edge in $G(n, m, p)$. For vertices $v, w \in V$,

$$P(vw) = 1 - P(S_v \cap S_w = \emptyset) = 1 - (1 - p^2)^m.$$

The following notation will be used throughout the paper in reference to the random models introduced in this section.

vw	the edge between vertices v and w
$\mathcal{A}$	a graph property (i.e., a set of graphs that is closed under isomorphism)
$G(n,p)$	the Erdős–Rényi independent-edge model of random graph
$G_{n,p}$	an instance drawn from $G(n,p)$
$G(n,M)$	the Erdős–Rényi fixed-size model of random graph
$G_{n,M}$	an instance drawn from $G(n,M)$
$R(n,m,p)$	the random matrix model with parameters n, m, and p (and the sample space of that model)
$R_{n,m,p}$	an instance drawn from $R(n,m,p)$
$G(n,m,p)$	the random intersection graph model with parameters n, m, and p (and the sample space of that model)
$G_{n,m,p}$	an instance drawn from $G(n,m,p)$
Ψ	the function that converts a matrix representation to a random intersection graph
$\{G_{n,m,p} \in \mathcal{A}\}$	the event that $G_{n,m,p}$ exhibits the property $\mathcal{A}$ (i.e., is an element of the graph set $\mathcal{A}$)

3 Random Graph Analysis

Much of the work that is done on the Erdős–Rényi $G(n,p)$ model seeks to characterize the random graph's "evolution", as $p = p(n)$ varies. Typically, it is assumed that $n \to \infty$, and that $p(n) \to 0$ as $n \to \infty$. The process of setting p equal to functions that are successively larger in an asymptotic

sense (they converge to 0 at a slower rate) is then described as "increasing p". As p is increased in this way, many graph-theoretic events for $G(n,p)$ exhibit rather sudden jumps in probability. For large n, the jump is usually from probability near 0 to probability near 1.

One of the main goals in analyzing the evolution of any random graph model is the determination of threshold functions for almost sure possession of a particular graph property. In this context, *almost sure* means that the limiting probability of the event is 1 as $n \to \infty$. (This contrasts with the probabilist's standard usage of the term.) The abbreviation *a.s.* will often be used for almost sure or almost surely. A threshold function $t(n)$ for graph property $\mathcal{A}$, as defined, for example, in [1], is a function such that

1. when $p(n) = o(t(n))$, $\lim_{n\to\infty} P(G \text{ has property } \mathcal{A}) = 0$, and
2. when $p(n) = \omega(t(n))$, $lim_{n\to\infty} P(G \text{ has property } \mathcal{A}) = 1$,

or vice versa. For example, the following two propositions describe thresholds for basic properties within the $G(n,m,p)$ model.

Proposition 1 *(***edge appearance threshold at** $1/(n\sqrt{m})$*) If* $p(n) = 1/(\omega n\sqrt{m})$ *for some* $\omega \to \infty$*, then* $G_{n,m,p}$ *a.s has no edges. If* $p(n) = \omega/(n\sqrt{m})$ *for some* $\omega \to \infty$*, then* $G_{n,m,p}$ *a.s. has at least one edge.*

Proposition 2 *(***complete graph threshold at** $\sqrt{(2\ln n)/m}$*) If* $p(n) = \sqrt{\frac{2\ln n - \omega}{m}}$ *for some* $\omega \to \infty$*, then* $G_{n,m,p}$ *a.s. has nonedges. If* $p(n) = \sqrt{\frac{2\ln n + \omega}{m}}$ *for some* $\omega \to \infty$*, then* $G_{n,m,p}$ *is a.s. a complete graph.*

The proofs of these propositions use the first and second moment methods; see [12]. They also follow directly from the more general subgraph Theorem 3 in [8]. In words, Propositions 1 and 2 show that when p is extremely small, the matrix $R_{n,m,p}$ is almost surely so sparse that the resulting graph $G_{n,m,p}$ has no edges. At the other end of the spectrum, when p is close enough to 1, the matrix is dense and $G_{n,m,p}$ is almost surely a complete graph. This report thus focuses on $p \to 0$ such that

$$\frac{\omega}{n\sqrt{m}} \leq p \leq \sqrt{\frac{2\ln n - \omega}{m}} \text{ for some } \omega \to \infty \text{ as } n \to \infty.$$

All such values of p satisfy $mp^3 \to 0$, since $mp^3 \leq m\left(\sqrt{\frac{2\ln n}{m}}\right)^3 = \frac{(2\ln n)^{3/2}}{m^{1/2}} = \frac{(2\ln n)^{3/2}}{n^{\alpha/2}} \to 0$.

Threshold functions for graph properties will form the basis of our comparison of $G(n,m,p)$ with $G(n,p)$. Below are some additional threshold results that have been found for the two models. The first results given are the subgraph thresholds (for containment of a particular kind of small graph as a subgraph). For $G(n,m,p)$ these are solely for induced subgraphs and follow from the more general subgraph Theorem 3 in [8], which requires a fair bit of notation and so is not given here. Instead, we state the thresholds for three specific induced subgraphs in Table 1.

When any of the functions given in Table 1 is used for p, the relation $mp^2 \to 0$ is forced. Thus approximation 1 (from the Introduction) implies that the asymptotic probability of an edge can be calculated for each threshold p function as mp^2. This edge probability is included in the third column of the table for later reference.

Table 1: $G(n, m, p)$: Appearance thresholds for some induced subgraphs

Graph	Threshold Function	$\sim$ Edge Probab. at Threshold
Z_h(cycle on $h \geq 4$ vertices)	$p = \frac{1}{n^{\frac{1}{2}} m^{\frac{1}{2}}}$	$\frac{1}{n}$
K_h(complete on $h \geq 2$ vertices)	$p = \begin{cases} \frac{1}{nm^{\frac{1}{h}}}, & \alpha \leq \frac{2h}{h-1} \\ \frac{1}{n^{\frac{1}{h-1}} m^{\frac{1}{2}}}, & \alpha \geq \frac{2h}{h-1} \end{cases}$	$\frac{m^{1-\frac{2}{h}}}{n^2}$ $\frac{1}{n^{\frac{2}{h-1}}}$
P_h(path on $h \geq 2$ vertices)	$p = \begin{cases} \frac{1}{n^{\frac{1}{2}} m^{\frac{h-1}{2(h-2)}}}, & \alpha \leq \frac{h-2}{h-1} \\ \frac{1}{n^{\frac{h}{2(h-1)}} m^{\frac{1}{2}}}, & \alpha \geq \frac{h-2}{h-1} \end{cases}$	$\frac{1}{nm^{\frac{1}{h-2}}}$ $\frac{1}{n^{\frac{h}{h-1}}}$

For $G(n, p)$ the threshold for containment of a subgraph H always exists and is simpler to describe. Using the notation in [7], where

- $d(H) = |E(H)|/|V(H)|$ is the edge density of H,
- $m(H) = \max\{d(L) : L \text{ is a subgraph of } H\}$ is the maximum subgraph density of H,

we have

Theorem 3 *The threshold for the property that $G(n,p)$ contains a copy of H is*

$$\frac{1}{n^{1/m(H)}}.$$

The $G(n,p)$ thresholds for the 3 examples that were described under the $G(n,m,p)$ model are given in Table 2.

Table 2: $G(n,p)$: Appearance thresholds for some subgraphs

Graph	Threshold Function τ_H
Z_h(cycle on $h \geq 4$ vertices)	$p = \frac{1}{n}$
K_h(complete on $h \geq 2$ vertices)	$p = \frac{1}{n^{2/(h-1)}}$
P_h(path on $h \geq 2$ vertices)	$p = \frac{1}{n^{h/(h-1)}}$

This is a good point at which to compare the $G(n,p)$ subgraph threshold p with the asymptotic edge probability at the corresponding subgraph thresholds in the $G(n,m,p)$ model (third column of Table 1). Some are the same and some are not. To illuminate a set of cases where these functions are certain to be the same it is important to compare Theorem 3 with the following theorem from Section 2.1 of [8], pointing specifically to the cases when the parameter α is large.

Theorem 4 *For $G(n,m,p)$ and any fixed graph H, there is an $\alpha^* > 0$*

such that for all $\alpha \geq \alpha^*$ *the induced subgraph appearance threshold for* H *is* $p = \frac{1}{n^{1/(2m(H))} m^{1/2}}$.

In Section 2.3.3 of [12] it is shown that $\alpha^* = 3$ is always big enough, i.e., for any $\alpha \geq 3$, the appearance threshold is as stated. The asymptotic edge probability at this threshold, computed as mp^2, is $1/(n^{1/m(H)})$, the same as the threshold given in Theorem 3.

Shifting attention to another property of interest, the threshold for connectivity in $G(n,p)$ is given in the following theorem.

Theorem 5 *Let* $p = c\frac{\ln n}{n}$.

- *If* $0 < c < 1$, *then* $G(n,p)$ *is a.s. disconnected.*
- *If* $c > 1$, *then* $G(n,p)$ *is a.s. connected.*

(See [11] for a description of this result.)

The counterpart connectivity threshold theorem for $G(n,m,p)$ describes a threshold called τ_c.

Theorem 6 *For the model* $G(n,m,p)$,

$$\tau_c = \begin{cases} \frac{\ln n}{m} & \text{when } \alpha \leq 1 \\ \sqrt{\frac{\ln n}{nm}} & \text{when } \alpha > 1. \end{cases}$$

In fact,

for $\alpha \leq 1$

- *if* $p = (\ln n - \omega)/m$ *with* $\omega \to \infty$*, then a.s.* $G_{n,m,p}$ *is disconnected, and*
- *if* $p = (\ln n + \omega)/m$ *with* $\omega \to \infty$*, then a.s.* $G_{n,m,p}$ *is connected,*

while for $\alpha > 1$

- *if* $p = \sqrt{(\ln n - \omega)/(nm)}$ *with* $\omega \to \infty$*, then a.s.* $G_{n,m,p}$ *is disconnected, and*
- *if* $p = \sqrt{(\ln n + \omega)/(nm)}$ *with* $\omega \to \infty$*, then a.s.* $G_{n,m,p}$ *is connected.*

(See Section 3.2 of [12] for proof.)

4 Comparison of Evolutions

How might the thresholds for the two different models be compared, other than to say whether the asymptotic edge probability at thresholds for certain properties is the same? Consider the evolution of $G(n, m, p)$ as p is increased through a range of various functions for which $mp^2 \to 0$, and that of $G(n, p)$ as p is increased. In which order are the different graph thresholds reached? Are the cliques the first fixed graphs to arrive or the last? The examples of the previous section showed that for $G(n, m, p)$ even the subgraph thresholds for one particular subgraph can differ greatly, depending on the size of α, and hence on the relationship between m and n. Not only is the threshold

usually a function of α, but some threshold functions exhibit different forms according to the range in which α falls. For complete graphs, for example, the threshold is $\frac{1}{nm^{1/h}}$ $\left(= \frac{1}{n^{1+(\alpha/h)}}\right)$ if α is less than $\frac{2h}{h-1}$ and $\frac{1}{n^{1/(h-1)}m^{1/2}}$ otherwise.

A timeline for the early evolution of $G(n, m, p)$ when $\alpha = 1/2$ (referenced as $G(n, n^{1/2}, p)$) and for $G(n, p)$ will serve to illustrate the analysis of threshold progression. The key for comparison here continues to be in examining the probability of an edge at each of the thresholds. Figure 2 shows the probability of an edge at the thresholds for triangles, h-cliques ($h \geq 4$), P_3, and cycles ($h \geq 4$) in the two cases. The edge probability at the connectivity threshold is also included for comparison. Note that the edge probability at the threshold for P_3 is the same in both cases. This is also true for the cycle threshold. However, the thresholds for appearances of cliques are at quite different "times" for $G(n, n^{1/2}, p)$ and for $G(n, p)$. In the $G(n, n^{1/2}, p)$ case, the threshold for triangles is before that for P_3, while in the $G(n, p)$ case triangles appear later in the evolution. In fact, for $G(n, n^{1/2}, p)$ the asymptotic edge probability for an edge vw at the threshold for general K_h ($h \geq 4$ fixed) is

$$P(vw \text{ at } \tau_{K_h,1}) \sim \frac{1}{n^{(3/2)+(1/h)}},$$

which is less than $\frac{1}{n^{3/2}}$, the asymptotic edge probability at the P_3 threshold. So all fixed cliques appear very early in the evolution of $G(n, n^{1/2}, p)$.

The $G(n, n^{1/2}, p)$ timeline is evidence that in a small-m scenario ($m = n^{1/2}$) for $G(n, m, p)$ it is more "difficult" to form an induced pair of adjacent edges without also forming a third edge between them (i.e., to form P_3)

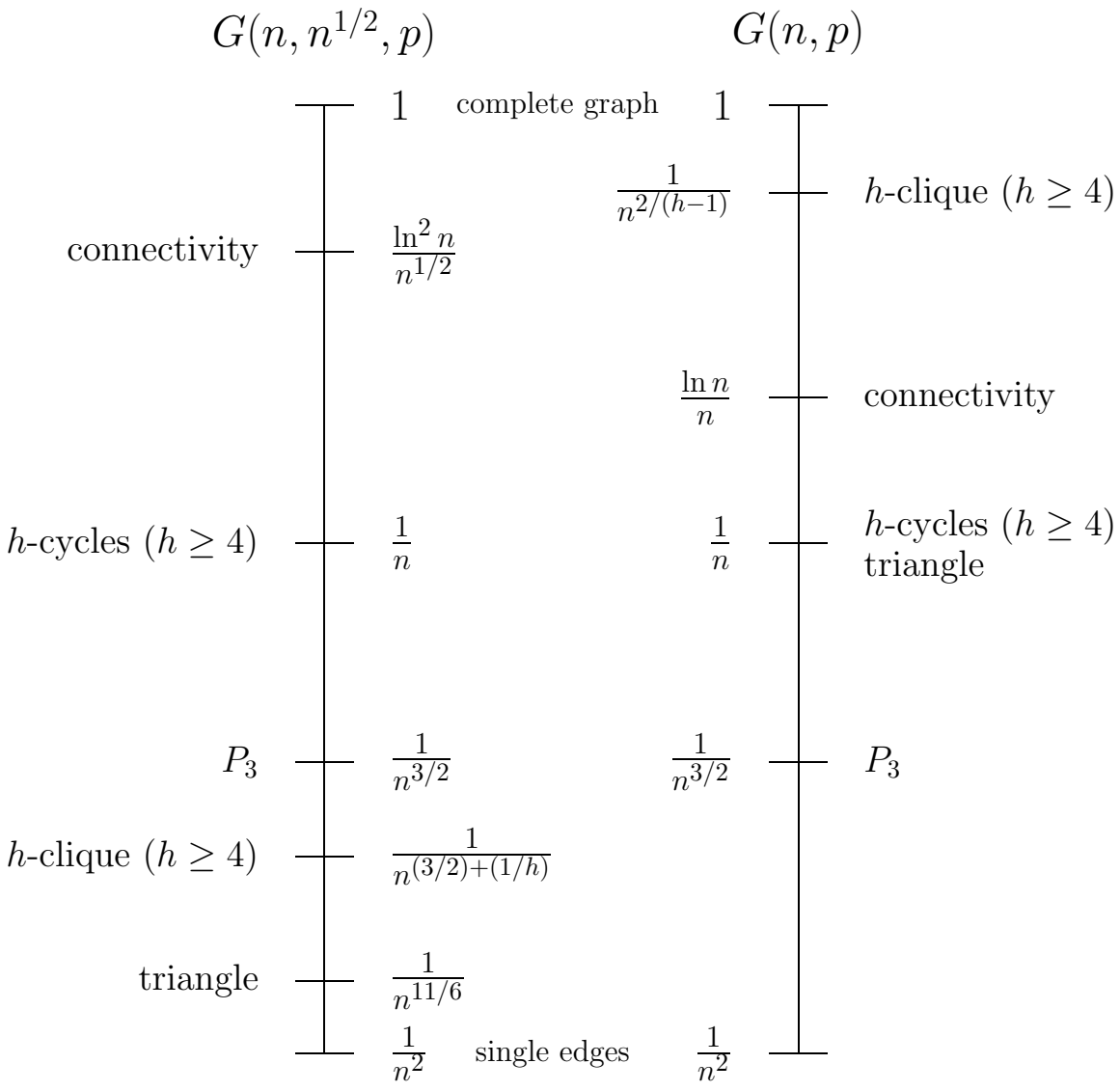

Figure 2: A comparison of (asymptotic) edge probability at various thresholds for $G(n, n^{1/2}, p)$ and $G(n, p)$

than to form a clique of any fixed size. The intersection representation for P_3 always requires the complementary activity of two column cliques, while a clique of fixed size can be generated by an individual column of the matrix $R_{n,m,p}$. When α is small, generation by individual columns is the more likely event, even for fairly small p. In $G(n,p)$, on the other hand, cliques occur much later in the evolution. Triangles have the same late threshold as cycles, and larger cliques do not have appearance thresholds until after the threshold for connectivity.

5 Introduction to Large-α Analysis

The preceding sections have illustrated the fact that for some values of α the evolution of $G(n,m,p)$ is quite different from that in the Erdős–Rényi model, $G(n,p)$. Yet for large values of α it seems to be quite similar. In fact, for the subgraph thresholds, when $\alpha \geq 3$, there is an equivalence with the threshold results for the Erdős–Rényi model, as indicated by comparing Theorem 4 for $G(n,m,p)$ with Theorem 3 for $G(n,p)$. (In this situation, as in the timelines of the previous section, equivalence refers to the fact that when the thresholds for appearance of subgraphs are expressed in terms of the probability of an edge at that threshold, the probabilities within the two models are asymptotically the same.) Furthermore, the large-α threshold for connectivity of $G_{n,m,p}$ is equivalent in this same sense to the Erdős–Rényi connectivity threshold, since for all $\alpha > 1$ the expected number of edges at the connectivity threshold is of order $\frac{1}{2}n \ln n$ in both models.

To what extent does this equivalence generalize to thresholds and probabilities for other graph properties? For what ranges of α do the models behave similarly? It is now time to explore these topics, in an effort to discover how the new $G(n, m, p)$ model relates to existing work on $G(n, p)$.

The case in which the exponent α is large corresponds to parameter choices with the number of universal set elements, m, very much larger than the number of vertices in the graph, n. While large α implies a relatively plentiful set of possible elements that can be picked by a vertex, the probability that the vertex picks any particular element is appropriately small. This balance is enforced in order that the probability p remain in the range of interest (described at the beginning of Section 3), where the graph realization is a.s. not edgeless or complete, i.e., in the range $1/(nm^{1/2}) \leq p \leq \sqrt{(2 \ln n)/m}$.

As a result, the probability for picking a particular element is so small that the occurrence of columns of the random matrix that have more than two 1's (i.e., elements that are picked by more than two vertices) has only insignificantly small probability. Thus the dependence between edge events[1] or between non-edge events is relatively small. When α is large the knowledge that two vertices pick a common element no longer holds much information about the likelihood that they share elements with additional vertices.

For example, consider the events of edges vw, wx, and vx, when $mp^2 \to$

[1]The term *edge event* refers here to an event of the form $\{e_{i_1}, e_{i_2}, \ldots, e_{i_k}\}$, where e_{i_j} means that edge e_{i_j} appears in the graph and where the presence or absence of additional edges is ignored.

0. These three edges, if they occur in the graph, form a triangle on vertices v, w, and x. Given that the random intersection graph contains edges vw and wx, what is the probability that the third edge vx exists, closing the triangle? By bound 1 of Section 1, the marginal probability of any particular edge is asymptotically $mp^2 \to 0$. The conditional probability compares with the marginal in different manners for the two extreme ranges of α.

When α is small enough, the conditional probability of a third edge is significantly greater than the marginal probability. With a comparatively small total number of elements in the universal set (or total number of column cliques), large cliques are the common way in which adjacent edges are generated. Under this scenario it is quite likely that the existing vw and wx were generated by a triangle vwx for a shared element, and thus that edge vx is in the graph.

For large α, however, the conditional probability of edge vx is still asymptotically mp^2, the magnitude of the unconditional edge probability. As described above, the probability that a vertex picks a particular element is so small that the matrix in this situation is not likely to create three adjacencies via a single column. Thus the probability that vx is generated in its own column is the dominating one. This edge column probability is virtually unaffected by the condition of existence of columns generating the other two edges.

For computing the conditional probabilities to illustrate these extreme cases, we denote the events of interest by

$$A \;=\; \{R_{n,m,p} \text{ has a column containing 1's for } v,\, w,\, x\}$$

$$
\begin{aligned}
B &= \{R_{n,m,p} \text{ has individual columns generating } vw,\, wx,\, vx\} \\
C &= \{R_{n,m,p} \text{ has individual columns generating } vw,\, wx\}.
\end{aligned}
$$

Since the edges can be generated in a triangle or individually in columns, the conditional probability can be written as

$$
P(vx \,|\, vw, wx) = \frac{P(\text{triangle } vwx)}{P(\text{edges } vw, wx)} = \frac{P(A \cup B)}{P(A \cup C)}.
$$

Theorem 2.3 in [12] (requiring notation not used here) implies that

$$
\begin{aligned}
P(A) &\sim mp^3 \\
P(B) &\sim (mp^2)^3 \\
P(C) &\sim (mp^2)^2.
\end{aligned}
$$

Using these expressions in some basic probability inequalities gives bounds for unions of the events, as

$$
\begin{array}{lllll}
P(A \cup B) & \geq & P(A) & \sim & mp^3 \\
P(A \cup B) & \geq & P(B) & \sim & (mp^2)^3 \\
P(A \cup B) & \leq & P(A) + P(B) & \sim & mp^3 + (mp^2)^3 \\
P(A \cup C) & \geq & P(C) & \sim & (mp^2)^2 \\
P(A \cup C) & \leq & P(A) + P(C) & \sim & mp^3 + (mp^2)^2.
\end{array}
$$

Consider first as an example the specific small-α case $\alpha = 1/4$, $p = 1/(m^{1/2}n^{1/4})$. This p is small enough so that $mp^2 = n^{-1/2} \to 0$, but large enough that the graph a.s. contains a triangle. Substitution of these values for α and p gives the result that $mp^3 = 1/n^{7/8}$, $(mp^2)^2 = 1/n$, and $(mp^2)^3 = 1/n^{3/2}$. Since mp^3 is clearly of greater order than the other terms, the asymptotic bounds above lead to the conclusion that the conditional probability is much higher than the marginal one:

$$
P(vx \,|\, vw, wx) \geq \frac{P(A)}{P(A) + P(C)} \sim \frac{mp^3}{mp^3 + (mp^2)^2} = 1 - o(1) \gg (n^{-1/2}) \sim (mp^2) \sim P(vx).
$$

Now consider instead any $\alpha > 6$. Comparing the terms $(mp^2)^3$ and (mp^3) shows that $(mp^2)^3 = (mp^3)(m^2p^3)$, where

$$m^2p^3 \geq \frac{m^2}{n^3m^{3/2}} = \frac{m^{1/2}}{n^3} \to \infty \text{ (using } p \geq \tfrac{1}{nm^{1/2}}).$$

Thus

$$\begin{aligned} (mp^2)^3 &\gg (mp^3), \text{and certainly} \\ (mp^2)^2 &\gg (mp^2)^3 \gg (mp^3). \end{aligned}$$

Hence

$$P(vx \,|\, vw, wx) = \frac{P(A \cup B)}{P(A \cup C)} \sim \frac{(mp^2)^3}{(mp^2)^2} = mp^2 \sim P(vx)$$

for all such cases. The main result of this paper (Theorem 10) proves equivalence between $G(n, m, p)$ and the Erdős–Rényi model of random graphs on n vertices, when the parameter m is large relative to n. Specifically, equivalence is proved for $\alpha > 6$. Some of the motivation for this proof came from [9].

Taking p to be the probability that a particular vertex picks a particular element in the $G(n, m, p)$ model, the analysis relates $G(n, m, p)$ to $G(n, \hat{p})$ via $G(n, M)$, where $\hat{p}$ is the probability of a particular edge for both $G(n, m, p)$ (asymptotic probability) and $G(n, \hat{p})$ (exact probability), and M is the number of edges for a fixed-size random graph model. Under these circumstances, with $\alpha > 6$, Theorem 10 implies that $G(n, m, p)$ and $G(n, \hat{p})$ have the same thresholds for all graph properties.

The theorem is stated in terms of total variation distance between the probability measures for the two models. It thus bounds the absolute difference between $P(G_{n,m,p} \in \mathcal{A})$ and $P(G_{n,\hat{p}} \in \mathcal{A})$ for any graph property

$\mathcal{A}$. In fact, it implies that if for some p the event $\{G$ exhibits the specified property$\}$ has probability tending to a limit in one of the models, then it tends to a limit in the other and the two limiting probabilities are the same. Section 7 describes some special cases (in terms of ranges for the function $p(n)$) for which the equivalence results can be extended even when $\alpha < 6$. The discussion also gives a counterexample showing that the theorem cannot be applied generally for any case when $\alpha < 3$.

6 Equivalence of Models for Large α

This section uses the following global assumption.

Assumption *The exponent α satisfies $\alpha > 6$.*

The proof of the equivalence theorem begins with a proposition identifying an event that is of very low probability under this assumption. This essentially allows the event to be removed from consideration in computing $P(G_{n,m,p} \in \mathcal{A})$. Results about total variation distance are then applied within the resulting probability space to show implication relationships between conditions that form a path to the theorem's conclusion. Once the links in the path are constructed, one of the initial conditions is shown to hold, implying the following ones and thus completing the proof. As usual, the analysis focuses on p in the range where $G_{n,m,p}$ is a.s. not empty and a.s. not complete, i.e., where $\frac{\omega}{n\sqrt{m}} \leq p \leq \sqrt{\frac{2\ln n - \omega}{m}}$ for some $\omega \to \infty$. The definition of *total variation distance* for random variables on a finite set is given below. Some standard facts about this quantity will be proved and

used subsequently in proving the theorem.

Suppose X and Y are random variables taking values in a common finite set S. Consider $\mathcal{L}(X)$ and $\mathcal{L}(Y)$, the probability measures on S whose values at $A \subseteq S$ are $P(X \in A)$ and $P(Y \in A)$, respectively.

Definition 1 With X, Y, $\mathcal{L}(X)$, and $\mathcal{L}(Y)$ as above, the *total variation distance* between $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ is defined as

$$\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) := \max_{A \subseteq S} |P(X \in A) - P(Y \in A)|.$$

Equivalently, the total variation distance can be written as

$$\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) = \frac{1}{2} \sum_{x \in S} |P(X = x) - P(Y = x)|.$$

Note that X and Y may be graph-valued random variables, with $S = G_n$. The total variation distance can thus provide a measure of comparison between random graph models.

The first proposition in this section describes a "reduced sample space" of representation matrices. Note that since $p \leq \sqrt{\frac{2 \ln n}{m}}$, where $\alpha > 6$ (so that $m = \omega(n^6)$), it is always the case that

- $np \leq \frac{\sqrt{2 \ln n}}{n^2} \to 0$
- $mn^3p^3 \leq mn^3(\frac{2 \ln n}{m})^{3/2} = \frac{2\sqrt{2}n^3(\ln n)^{3/2}}{m^{1/2}} \to 0.$

Proposition 7 *Let T be the event that $R_{n,m,p}$ has at least one column containing three or more 1's (a "triple"). Then $P(T) = O(mn^3p^3) = o(1)$.*

Proof. Let X be the number of triples of ones that occur in columns of R, where a cell may be counted in multiple triples. Then T occurs if and only if $X > 0$. Since $\mathbf{E}X = m\binom{n}{3}p^3 \sim (mn^3p^3)/6$, Markov's inequality (the first moment method) implies

$$P(T) = P(X > 0) \leq \mathbf{E}X \sim (mn^3p^3)/6 \to 0.2$$

This proposition implies that uniformly for any graph event $\mathcal{A}$,

$$P(G \in \mathcal{A}) = P(G \in \mathcal{A}|T^c)(1-O(mn^3p^3))+O(mn^3p^3) = P(G \in \mathcal{A} \,|\, T^c)+o(1).$$

The sequel is thus concerned only with probabilities of events conditioned on T^c, the event that all columns of R have at most two 1's. For convenience, the following notation is defined.

Definition 2 The probability of an event A, conditioned on the event that $R_{n,m,p}$ contains no triples, will be written as $P_{T^c}(A)$.

Using this notation, the result of Proposition 7 can be expressed as

$$\mathrm{d}(\mathcal{L}(G_{n,m,p}), \mathcal{L}_{T^c}(G_{n,m,p})) \to 0.$$

It is now possible to compute the asymptotic probability for picking a particular edge e in the model $G(n, m, p)$ when $\alpha > 6$. Note that columns are still independent, conditionally given T^c. Considering a particular column of the representation matrix, p_e is defined by

Definition 3 $p_e := P_{T^c}(\text{edge } e \text{ is generated by a particular column}).$

It is straightforward to compute that

Lemma 8

$$p_e = \frac{p^2}{q^2 + npq + \binom{n}{2}p^2}.$$

Proof.

$$\begin{aligned} p_e &= \frac{P(\text{the column has exactly the two 1's needed and no third 1})}{P(\text{the column has no triples})} \\ &= \frac{p^2 q^{n-2}}{q^n + npq^{n-1} + \binom{n}{2}p^2 q^{n-2}} \\ &= \frac{p^2}{q^2 + npq + \binom{n}{2}p^2}. \end{aligned}$$

Given any $\epsilon_1 > 0$ and $\epsilon_2 > 0$, the facts that $q \to 1$ and $np \to 0$ imply that n may be taken sufficiently large so that $q > 1 - \epsilon_1$ and $np < \epsilon_2$. The resulting inequalities

$$\left[q^2 + npq + \binom{n}{2}p^2\right] > 1 - 2\epsilon_1$$

and

$$\left[q^2 + npq + \binom{n}{2}p^2\right] < 1 + \epsilon_2 + \epsilon_2^2$$

imply that $q^2 + npq + \binom{n}{2}p^2$ can be made arbitrarily close to 1, i.e., that $q^2 + npq + \frac{(np)^2}{2} \sim 1$. Thus, $p_e \sim p^2$. Since the columns are conditionally independent and p_e is the same for all of them, the conditional probability of a particular edge can be computed exactly in terms of p_e. For $\hat{p}$ defined by

Definition 4 $\hat{p} := P_{T^c}(\text{edge } e \text{ in } G_{n,m,p})$,

this probability is given by

Lemma 9

$$\hat{p} = 1 - (1 - p_e)^m.$$

This is derived exactly as is the standard edge probability in $G(n, m, p)$, except that p_e is substituted for p^2 in compliance with the condition of no triples. Note that $\hat{p}$ is the same for each of the $\binom{n}{2}$ possible edges. It is now time for the main theorem.

Theorem 10 (Model Equivalence for Large α) *Let $m = n^\alpha$ and $\alpha > 6$. Let $p = p(n)$ be such that $\frac{\omega}{n\sqrt{m}} \leq p \leq \sqrt{\frac{2 \ln n - \omega}{m}}$ for some $\omega \to \infty$ (so that the graph is a.s. not edgeless and not complete). Let $\hat{p}$ be as defined above. Then*

$$\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}(G_{n,m,p})) \to 0 \text{ as } n \to \infty.$$

Corollary 11 *Under the conditions above and for any $a \in [0, 1]$ and any graph property $\mathcal{A}$, as $n \to \infty$ it follows that*

$$P(G_{n,m,p} \in \mathcal{A}) \to a \text{ iff } P(G_{n,\hat{p}} \in \mathcal{A}) \to a.$$

In particular,

$$P(G_{n,m,p} \in \mathcal{A}) \to 1 \text{ iff } P(G_{n,\hat{p}} \in \mathcal{A}) \to 1$$

and

$$P(G_{n,m,p} \in \mathcal{A}) \to 0 \text{ iff } P(G_{n,\hat{p}} \in \mathcal{A}) \to 0.$$

The corollary implies that when $\alpha > 6$, the two random graph models have the same event-threshold probability functions, where the thresholds are expressed in terms of the probability of an edge in the respective models.

Proof. (of Theorem 10) The following two random variables are of key importance in this proof. For convenience, define $N := \binom{n}{2}$. Let

$$Y \sim \text{Bin}(N, \hat{p}) \text{ and } X = \sum_{i=1}^{N} X_i, \tag{1}$$

where $X_1, \ldots, X_N$ are indicator random variables for the slightly dependent events in the following scheme. Imagine a game in which $(N+1)$ possible coupons $c_1, \ldots, c_{N+1}$ are available. The player makes m draws, with replacement, such that on any particular draw,

$$\begin{array}{rlll} P(c_i) = & p_e, & i = 1, \ldots, N & \text{(winning coupons)} \\ P(c_{N+1}) = & 1 - Np_e & & \text{(blank coupon)} \end{array}$$

The assumptions and definition of p_e, which imply that $p_e \sim p^2$, then imply too that $Np_e \asymp (np)^2 \to 0$. Now, let

$$X_i = \begin{cases} 1 \text{ if } c_i \text{ is picked in at least one of the } m \text{ draws} \\ 0 \text{ otherwise.} \end{cases}$$

Thus $\mathbf{E}X_i = 1-(1-p_e)^m = \hat{p}$ and $\mathbf{E}X = N\hat{p} = \mathbf{E}Y$. The random variable X is the number of different types of winning coupons picked in the m draws. This corresponds exactly to the random variable for the number of edges when probabilities in $G(n, m, p)$ are conditioned on T^c, i.e.,

Proposition 12

$$P_{T^c}(|E(G_{n,m,p})| = M) \equiv P(X = M).$$

What happens to $P(G_{n,m,p} \in \mathcal{A})$ under the conditions T^c and $\{|E(G_{n,m,p})| = M\}$, for given fixed M? It is then equal to the sum of the conditional probabilities (conditioned on T^C) of those graphs with M edges that satisfy $\mathcal{A}$, divided by the sum of the conditional probabilities for all graphs with M edges. Under T^c, each graph with M edges is equally likely. Thus the probability of a particular subset of these graphs is the same as in the fixed edge-size model $G(n, M)$. This implies the useful equality that

Proposition 13

$$P_{T^c}(G_{n,m,p} \in \mathcal{A} \Big| \, |E(G_{n,m,p})| = M) \equiv P(G_{n,M} \in \mathcal{A}).$$

In order to simplify the proof that $\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}(G_{n,m,p})) \to 0$, the following facts about total variation distance are introduced. As in the definition given above for total variation distance, the random variables X and Y are assumed to take values on a finite set S. For more information on total variation distance, see, e.g., [3].

Fact 1. Total variation distance is a metric on the space of all probability measures on S. (This is easily checked.)

Fact 2. If Z is a random variable on the same space as Y, then an upper bound on $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y))$ is given by

$$\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \leq \sum_z P(Z = z)\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y|Z = z)).$$

Proof. For any $A \subseteq S$,

$$|P(X \in A) - P(Y \in A)| \;\; = \;\; \left| P(X \in A) - \sum_z P(Z = z)P(Y \in A|Z = z) \right|$$

$$
\begin{aligned}
&= \left| \sum_z P(Z=z)[P(X \in A) - P(Y \in A | Z=z)] \right| \\
&\leq \sum_z P(Z=z) \left| P(X \in A) - P(Y \in A | Z=z) \right| \\
&\leq \sum_z P(Z=z) \mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y|Z=z)).
\end{aligned}
$$

The desired inequality thus follows when the maximum is taken over all A.

2

Fact 3. Suppose that there exist random variables Z and Z' such that

$$\mathcal{L}(X|Z=z) = \mathcal{L}(Y|Z'=z) \text{ for all } z.$$

Then

$$\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \leq 2\mathrm{d}(\mathcal{L}(Z), \mathcal{L}(Z')).$$

Proof.

$$
\begin{aligned}
|P(X \in A) - P(Y \in A)| &= \left| \sum_z P(Z=z)P(X \in A|Z=z) - \sum_z P(Z'=z)P(Y \in A|Z'=z) \right| \\
&= \left| \sum_z [P(Z=z) - P(Z'=z)]P(X \in A|Z=z) \right| \\
&\leq \sum_z |P(Z=z) - P(Z'=z)| P(X \in A|Z=z) \\
&\leq \sum_z |P(Z=z) - P(Z'=z)| \\
&= 2\mathrm{d}(\mathcal{L}(Z), \mathcal{L}(Z')).2
\end{aligned}
$$

It is not hard to improve the constant from 2 to 1: see p. 21 of [3].

Fact 4. If there exists a probability space on which random variables X' and Y' are both defined, with $\mathcal{L}(X) = \mathcal{L}(X')$ and $\mathcal{L}(Y) = \mathcal{L}(Y')$, then

$$\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \leq P(X' \neq Y').$$

Proof. For any $A \subseteq S$,

$$\begin{aligned} P(X \in A) - P(Y \in A) &= P(X' \in A) - P(Y' \in A) \\ &= [P(Y' \in A, X' = Y') + P(X' \in A, X' \neq Y')] - P(Y' \in A) \\ &\leq P(Y' \in A) + P(X' \neq Y') - P(Y' \in A) \\ &= P(X' \neq Y'). \end{aligned}$$

Reversing the roles of X and Y, we conclude

$$|P(X \in A) - P(Y \in A)| \leq P(X' \neq Y').2$$

It is now possible to reduce the problem of showing that $\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}(G_{n,m,p})) = o(1)$ to showing that $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \to 0$, where X and Y are the random variables defined in (1). This is accomplished as follows. Since Proposition 7 shows that

$$\mathrm{d}(\mathcal{L}(G_{n,m,p}), \mathcal{L}_{T^c}(G_{n,m,p})) \to 0,$$

Fact 1 and the triangle inequality imply that it is sufficient to show that

$$\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}_{T^c}(G_{n,m,p})) \to 0.$$

But Proposition 13 says that

$$\mathcal{L}_{T^c}(G_{n,m,p} \Big| \ |E(G_{n,m,p})| = M) = \mathcal{L}(G_{n,M}),$$

and it is clear that

$$\mathcal{L}(G_{n,\hat{p}} \Big| \ |E(G_{n,\hat{p}})| = M) = \mathcal{L}(G_{n,M}),$$

since all graphs with M edges have equal weight in this conditional measure. Thus by Fact 3 it is enough to show that

$$\mathrm{d}(\mathcal{L}(|E(G_{n,\hat{p}}|), \mathcal{L}_{T^c}(|E(G_{n,m,p})|)) \to 0.$$

Furthermore, since

$$\mathcal{L}(|E(G_{n,\hat{p}}|) = \mathrm{Bin}(N, \hat{p}) = \mathcal{L}(Y)$$

and

$$\mathcal{L}_{T^c}(|E(G_{n,m,p})|) = \mathcal{L}(X) \text{ (by Proposition 12)}$$

it suffices to show that $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \to 0$. Henceforth, the proof shall be concerned only with justifying the claim that

Claim 1 $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \to 0$.

Proof of Claim 1. In lieu of direct comparison between the laws of X and Y, a third random variable, $X(M)$, is defined and its probability distribution is compared with the others. Applying the triangle inequality to bounds on $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(X(M)))$ and $\mathrm{d}(\mathcal{L}(Y), \mathcal{L}(X(M)))$ then yields a bound on $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y))$. The new random variable $X(M)$ is defined with respect to another random variable, $M \sim \mathrm{Poi}(m)$, as follows.

> *Modify the coupon-picking scheme that defined the original random variable X by choosing the number of coupons to be picked according to the random M instead of the fixed m.*

Note that by definition of M as $M \sim \mathrm{Poi}(m)$, the expected number of picks is still m. As with X_i, define

$$X_i(M) = \begin{cases} 1 \text{ if } c_i \text{ is picked in at least one of the draws} \\ 0 \text{ otherwise.} \end{cases}$$

and let $X(M) = \sum_{i=1}^{N} X_i(M)$. In this probability model, the random variables $R_1, R_2, \ldots, R_N$ representing the numbers of draws of the different coupon types are independent Poisson random variables with mean mp_e.

Their distributions are verified by examination of their joint probability mass function, which is separable into a product of Poisson mass functions. For $i = 1, 2, 3, \ldots, N$, the probability that there is at least one draw of type i is thus given by

$$P(X_i(M) = 1) = 1 - e^{-mp_e},$$

and $P(X_i(M) = 0) = e^{-mp_e}$. Moreover, since $X_1, \ldots, X_N$ are mutually independent, the distribution of the sum $X(M)$ is $X(M) \sim \text{Bin}(N, 1 - e^{-mp_e})$. A coupling of the binomials Y and $X(M)$ is achieved by considering N independent uniform random variables $V_1, V_2, \ldots V_N$ and setting

$$\begin{aligned} Y_i = 1 \quad &\text{if and only if} \quad V_i \leq 1 - (1 - p_e)^m, \text{ and} \\ X_i(M) = 1 \quad &\text{if and only if} \quad V_i \leq 1 - e^{-mp_e}. \end{aligned}$$

Since $1 - (1 - p_e)^m \geq 1 - e^{-mp_e}$, this coupling ensures that $Y \geq X(M)$, so that

$$\begin{aligned} P(Y \neq X(M)) &= P(\text{at least one of the uniform random variables} \\ &\quad \text{falls in the interval } (1 - e^{-mp_e}, 1 - (1 - p_e)^m]) \\ &\leq N \cdot P(V_i \in (1 - e^{-mp_e}, 1 - (1 - p_e)^m]) \\ &= N(e^{-mp_e} - (1 - p_e)^m) \\ &\leq Ne^{-mp_e} \left(\frac{mp_e^2}{2(1 - p_e)^2} \right) \text{ (see bound 3 of Section 1)} \\ &\leq Nmp_e^2 \\ &\ll n^2 m \left(\frac{\ln^2 n}{m^2} \right) \to 0. \end{aligned}$$

By Fact 4, this implies that $\text{d}(\mathcal{L}(Y), \mathcal{L}(X(M))) \to 0$ since

$$\text{d}(\mathcal{L}(Y), \mathcal{L}(X(M))) \leq P(Y \neq X(M)) \to 0.$$

Since the next goal is to compare X with $X(M)$, the random variable X will sometimes be written as $X(m)$. By Fact 2,

$$\begin{aligned} \mathrm{d}(\mathcal{L}(X(m)), \mathcal{L}(X(M))) &\leq \sum_r P(M=r)\mathrm{d}(\mathcal{L}(X(m)), \mathcal{L}(X(M)|M=r)) \\ &= \sum_r P(M=r)\mathrm{d}(\mathcal{L}(X(m)), \mathcal{L}(X(r))) \qquad (2) \\ &\quad (\text{ by def. of } X(M)). \end{aligned}$$

To see how to compare $\mathcal{L}(X(m))$ and $\mathcal{L}(X(r))$, consider more generally the integers $0 \leq s \leq s' < \infty$. Coupling $X(s')$ and $X(s)$, so that after a total of s' picks all of the picks are considered in determining $X(s')$ and only the first s picks for $X(s)$, yields

$$\begin{aligned} \mathrm{d}(\mathcal{L}(X(s')), \mathcal{L}(X(s))) &\leq P(X(s) \neq X(s')) \text{ (by Fact 4)} \\ &\leq P(X(s'-s) \geq 1) \\ &= 1-(1-Np_e)^{s'-s} \text{ (by Fact 4).} \end{aligned}$$

Thus $\mathrm{d}(\mathcal{L}(X(m)), \mathcal{L}(X(r))) \leq 1-(1-Np_e)^{|m-r|}$. From Chebyshev's inequality it follows that

$$P(|M-m| < m^{1/2}\ln m) \geq 1 - \frac{1}{\ln^2 m}.$$

Using this and the inequality in (3), the total variation distance $\mathrm{d}(\mathcal{L}(X(m)), \mathcal{L}(X(M)))$ can be bounded as follows:

$$
\begin{aligned}
\mathrm{d}(\mathcal{L}(X(m)), \mathcal{L}(X(M))) &\leq P(|M-m| \geq m^{1/2} \ln m) \\
&\quad + \sum_{r:|r-m|<m^{1/2}\ln m} P(M=r)[1-(1-Np_e)^{|m-r|}] \\
&\leq [P(|M-m| \geq m^{1/2}\ln m)\cdot 1 + P(|M-m| < m^{1/2}\ln m)\cdot 1] \\
&\quad - \sum_{r:|r-m|<m^{1/2}\ln m} P(M=r)(1-Np_e)^{|m-r|} \\
&= 1 - \sum_{r:|r-m|<m^{1/2}\ln m} P(M=r)(1-Np_e)^{|r-m|} \\
&\leq 1 - P(|M-m| < m^{1/2}\ln m)(1-Np_e)^{m^{1/2}\ln m} \\
&\leq 1 - \left(1 - \frac{1}{\ln^2 m}\right)(1-Np_e)^{m^{1/2}\ln m} \\
&\leq 1 - (1-o(1))(1 - Np_e m^{1/2}\ln m) \text{ by bound 1 of Section 1} \\
&\quad (\text{since } Np_e m^{1/2}\ln m \leq \frac{2n^2(\ln n)(\ln m)}{m^{1/2}} \to 0 \text{ for large } n) \\
&= o(1).
\end{aligned}
$$

Since the sufficient condition $\mathrm{d}(\mathcal{L}(X), \mathcal{L}(Y)) \to 0$ has been verified, the desired result that

$$\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}(G_{n,m,p})) \to 0$$

is proved. 2

7 Extending the Results: How Large is Large α?

The proof of the equivalence theorem (Theorem 10) of the previous section can be modified to give some equivalence results when α is not so large (i.e., for $\alpha < 6$), as long as p is small relative to m. In these cases, the random matrix still has a.s. no triple in a column, and the remaining part of the proof still holds. Specifically, the theorem's result can be extended to the following cases:

1. If $3 < \alpha \leq 6$ and $p \leq \sqrt{\frac{\ln n}{nm}}$ (p is below the connectivity threshold), then

   $$\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}(G_{n,m,p})) = o(1).$$

   The lower bound of 3 for α is chosen so that the expected number of triples, which may be as much as

   $$\frac{m\binom{n}{3}(\ln n)^{3/2}}{(mn)^{3/2}} \asymp \frac{(\ln n)^{3/2}}{n^{(\alpha-3)/2}},$$

   will have limit 0.

2. If $\alpha > 1.5$, and $mp \to 0$, then

   $$\mathrm{d}(\mathcal{L}(G_{n,\hat{p}}), \mathcal{L}(G_{n,m,p})) = o(1).$$

   The bound of 1.5 is chosen here to guarantee that the expected number of triple columns tends to 0. The expected number of such triples is of order

   $$mn^3p^3 = (mp)n^3p^2 \ll n^3p^2 = m^{3/\alpha}p^2.$$

   If $\alpha > 1.5$ then this tends to 0 since $mp \to 0$.

As noted in Theorem 4 the property of containing a fixed subgraph H as an induced subgraph has an equivalence between model thresholds for all $\alpha \geq 3$. Some of these thresholds are in the probability ranges for p given in items 1 and 2 above, while others are not. The equivalence for the wide range of properties encompassed by the subgraph theorem suggests that perhaps the main equivalence theorem can be extended for *all* properties when $3 \leq \alpha \leq 6$.

This strong model equivalence does not hold for the model when $\alpha < 3$. Specifically, for $p = 1/(n^{\frac{3}{4}} m^{\frac{5}{12}})$ and any $\alpha < 3$, there is a.s. a triangle in $G_{n,m,p}$ and a.s. no triangle in $G_{n,\hat{p}}$. This result is an application of the complete graph subgraph results from Tables 1 and 2. So for such α the existence of a triangle in the random graph is a property that does not always hold almost surely in one model exactly when it holds almost surely in the other. Further investigation of the differences between the models in this area would be of interest.

References

[1] N. Alon and J.H. Spencer, *The Probabilistic Method*, Wiley, New York, 1992.

[2] B. Bollobás, *Random Graphs*, Academic Press, Inc., New York, 1985.

[3] P. Diaconis, *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Haywood, CA, 1988.

[4] P. Erdős and A. Rényi, On random graphs I, *Publ. Math. Debrecen*, 6, 290–297 (1959).

[5] P. Erdős and A. Rényi, On the evolution of random graphs, *MTA Mat. Kut. Int. Kozl.*, 5, 17–61 (1960).

[6] T. Hagerup and C. Rub, A guided tour of Chernoff bounds, *Information Processing Letters*, 33, 305–308 (1990).

[7] M. Karoński, Random graphs. In R. L. Graham, M. Grotschel, and L. Lovasz, editors, *Handbook of Combinatorics*.

[8] M. Karoński, E.R. Scheinerman, K. Singer-Cohen, On random intersection graphs: the subgraph problem, *Combin. Probab. Comput.*, 8 (1999).

[9] T. Luczak, On the equivalence of two basic models of random graphs, *Random Graphs '87: Proceedings of the 3rd International Seminar on Random Graphs and Probabilistic Methods in Combinatorics*, 1990, pp. 151–157.

[10] E. Marczewski, Sur deux propriétés des classes d'ensembles, *Fund. Math.*, 33, 303–307 (1945).

[11] E.M. Palmer, *Graphical Evolution*, Wiley, New York, 1985.

[12] K. Singer, *Random Intersection Graphs*, Ph.D. thesis, The Johns Hopkins University, 1995.