Improved Random Graph Isomorphism

Tomek Czajka* Gopal Pandurangan*

Abstract

Canonical labeling of a graph consists of assigning a unique label to each vertex such that the labels are invariant under isomorphism. Such a labeling can be used to solve the graph isomorphism problem. We give a simple, linear time, high probability algorithm for the canonical labeling of a $G(n,p)$ random graph for $p \in \left[\omega\left(\ln^4 n/n \ln\ln n\right), 1-\omega\left(\ln^4 n/n \ln\ln n\right)\right]$. Our result covers a gap in the range of p in which no algorithm was known to work with high probability. Together with a previous result by Bollobas, the random graph isomorphism problem can be solved efficiently for $p \in [\Theta\left(\ln n/n\right), 1-\Theta\left(\ln n/n\right)]$.

1 Introduction

Random graph isomorphism is a classic problem in the algorithmic theory of random graphs [6, 4, 1]. In this problem, we are given a random $G(n,p)$ (Erdös-Renyi) graph and another graph H. The graph isomorphism problem is to decide whether the two graphs are isomorphic and if so, find an isomorphism between them. An isomorphism is a one-to-one mapping of vertices of G onto the vertices of H such that the edges of G are mapped onto the edges of H. Graph isomorphism can be solved by a *canonical labeling* of a graph [1, 2, 3]. Canonical labeling of a graph consists of assigning a unique label to each vertex such that the labels are invariant under isomorphism. More formally, given a class $\mathcal{K}$ of graphs which is closed under isomorphism, a canonical labeling algorithm assigns the numbers $1,\ldots,n$ to the vertices of each graph in $\mathcal{K}$, having n vertices, in such a way that two graphs in $\mathcal{K}$ are isomorphic if and only if the obtained labeled graphs coincide. The graph isomorphism problem can be solved using a canonical labeling of a graph as follows (e.g., [2]). Given a canonical labeling algorithm for $\mathcal{K}$, and an algorithm deciding whether a given graph belongs to $\mathcal{K}$ or not, we also have an algorithm deciding whether X is isomorphic to Y for any two graphs X, Y provided $X \in \mathcal{K}$. Namely, if $Y \notin \mathcal{K}$ then X is not isomorphic to Y; and if $Y \in \mathcal{K}$ then we have to check whether X and Y coincide after canonical labeling.

The first canonical labeling algorithm for random graphs was given by Babai, Erdös, and Selkow [1]. They gave a simple $O(n^2)$ time (linear in the number of edges) algorithm for

*Department of Computer Science, Purdue University, 250 N. Univ St., West Lafayette, IN 47907, USA. E-mail: {czajkat, gopal}@cs.purdue.edu.

canonical labeling of $G(n, 1/2)$ graphs with probability of failure bounded by $O(n^{-1/7})$. Since the $G(n, 1/2)$ model assigns a uniform distribution over all graphs (a total of $2^{\binom{n}{2}}$ graphs) the above result can be interpreted as an algorithm that succeeds on "almost all" graphs. This result was strengthened by Karp [7], Lipton [9], and Babai and Kucera [2]. In particular, Babai and Kucera [2] give a canonical labeling algorithm for the $G(n, 1/2)$ model that also runs in $O(n^2)$ time with exponential ($O(c^{-n})$) probability of rejection (i.e., not belonging to the canonical labeling class). In addition, they show that the rejected graphs can be handled such as to obtain a canonical labeling algorithm of all graphs with linear expected time, i.e., the average running time over the $2^{\binom{n}{2}}$ graphs is $O(n^2)$.

The question that motivates this paper and the line of research discussed next is this: Can we show a high probability canonical labeling algorithm for all p ($p = p(n)$)? Previous results have established this for various ranges of p. Bollobas [4, Theorem 3.17, page 74] gives a high probability linear time canonical labeling algorithm for $G(n, p)$, for $p = \omega(n^{-1/5} \ln n)$ and $p \leq 1/2$, i.e., for $p \in \left[\omega(n^{-1/5} \ln n), 1/2\right]$. (Note that for $p \geq 1/2$ one can equivalently consider the complement graph). The probability of algorithm failure is $O(n^{-1})$. We note that the above algorithm as well as the algorithms on $G(n, 1/2)$ cited earlier [1, 7, 9] all exploit properties of the degree sequence of a random graph. Another result of Bollobas [5] shows that canonical labeling can be done efficiently on much sparser graphs, i.e., for $\Theta\left(\frac{\ln n}{n}\right) \leq p \leq \Theta\left(n^{-11/12}\right)$. This result uses properties of the *distance sequence* of a vertex of a graph. The distance sequence of a vertex x is the list $\{d_i(x), 1 \leq i \leq n\}$ where $d_i(x)$ is the number of vertices at distance i from x. This algorithm takes $O(pn^3)$ time since all pairs of distances have to be computed. It is also known that if $0 \leq p \leq o(n^{-3/2})$ then the isomorphism problem is trivial [4] with high probability. To summarize, the ranges of p in which canonical labeling (and hence isomorphism) has been solved with high probability in polynomial time is:

$$\left[0, o\left(n^{-3/2}\right)\right], \left[\Theta\left(\frac{\ln n}{n}\right), o\left(\frac{1}{n^{11/12}}\right)\right], \left[\omega\left(n^{-1/5} \ln n\right), 1/2\right] \tag{1}$$

For each range we have an algorithm with polynomially small failure probability ($O(n^{-c})$ for some constant $c > 0$). For $p = 1/2$ the failure probability is exponentially small ($O(c^n)$ for some $0 < c < 1$).

This paper covers the gap between the last two ranges. We show a linear time, high probability canonical labeling algorithm for $G(n, p)$ graphs for $p = \omega(\ln^4 n/n \ln \ln n)$ and $p \leq 1/2$. Here, high probability will mean probability at least $1 - O(n^{-\alpha})$ for *every* constant $\alpha > 0$.

Our result significantly extends the range of p compared to [4, Theorem 3.17, page 74] and covers the gap between the second and third interval in (1). Our algorithm is similar to the Procedure A in [2], but simpler. However, they analyze the algorithm only for the $G(n, 1/2)$ random graph model. Our analysis is different from [2] and applies for a much larger range of p.

Our analysis uses an edge exposure martingale to analyze, given two vertices, how the degrees of their neighbors change as edges are added. This approach allows us to establish

good bounds on the probability of the two degree neighborhoods being the same, for a wide range of p.

2 The Algorithm

The idea of the algorithm is to distinguish all vertices of a graph using the degrees of their neighbors. We prove that this allows us to distinguish all vertices of a $G(n,p)$ graph (for sufficiently large p) with high probability. Define the *degree neighborhood* of a vertex as a sorted list of the degrees of the vertex's neighbors. We note that the degree and hence also the degree neighborhood are invariants under isomorphism. We use the degree neighborhood list to assign our canonical labeling, i.e., the label of a vertex is its degree neighborhood list.

The canonical labeling algorithm is as follows. It takes as input a graph G. The algorithm tries to assign a canonical labeling to the vertices of G by computing the degree neighborhood of each vertex. If the degree neighborhoods are not distinct the algorithm fails. To check for isomorphism, we can repeat the same procedure for H and then check whether the edges of G and H are same under the labelings.

1. Compute vertex degrees.
2. Compute degree neighborhoods for each vertex.
3. Sort vertices by degree neighborhoods in lexicographical order.
4. If the degree neighborhoods are not distinct for each vertex, FAIL.
5. Number the vertices in the sorted order.

Theorem 2.1 *If the algorithm does not fail, it outputs a canonical labeling of G.*
Proof: Steps 1 and 2 are invariant under isomorphism. If the algorithm does not fail, the computed degree neighborhoods are all distinct, hence steps 3 and 5 also are isomorphism invariant. Therefore the computed labeling is a canonical labeling. □

Theorem 2.2 *The algorithm can be implemented in linear time in graph's size ($O(V+E)$).*
Proof: Computing vertex degrees can clearly be done in linear time. Step 2 can be done in linear time by sorting all pairs (vertex, neighbor degree) in lexicographical order using radix-sort [8]. Step 3 is string sorting over the alphabet $\{0,\ldots,n-1\}$, this can be done in linear time [8]. □

Once we have a linear time canonical labeling algorithm, we can test for graph isomorphism in linear time: just compute the canonical labeling for both graphs G and H. Suppose the algorithm succeeds for G (we will prove this happens with high probability). If the algorithm fails for H, the graphs are not isomorphic. If it succeeds for H, sort the edges of both graphs lexicographically by the labels of their endpoints (using radix-sort) and compare the lists.

3 Failure Probability Analysis

Before we can analyze the algorithm, we need some preliminary lemmas on probability bounds concerning the binomial distribution.

p and other variables appearing in the proofs are functions of n. We will assume throughout the rest of the paper that $0 < p \leq 1/2$. All asymptotic notations such as $O(pn) = O(p(n)n)$ are taken as $n \to \infty$.

Let $B(n,p)$ denote the binomial distribution and $b(k;n,p) = \Pr(B(n,p) = k)$. Thus:

$$b(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Lemma 3.1 *If $pn = \Omega(1)$ then $b(k;n,p)$ is maximum for $k = \lfloor np \rfloor$ or $k = \lceil np \rceil$ and*

$$\max_k b(k;n,p) = \Theta\left(\frac{1}{\sqrt{pn}}\right)$$

Proof: See [4]. The formula follows from Stirling's approximation, in fact for $p = \omega\left(n^{-1}\right)$:

$$\max_k b(k;n,p) \approx \frac{1}{\sqrt{2\pi p(1-p)n}}$$

□

Lemma 3.2 *If $p = \omega(n^{-1})$ then:*

$$|b(k;n,p) - b(k-1;n-1,p)| = O\left(\frac{\sqrt{\ln n}}{pn}\right)$$

$$|b(k;n,p) - b(k;n-1,p)| = O\left(\frac{\sqrt{\ln n}}{pn}\right)$$

$$|b(k;n-1,p) - b(k-1;n-1,p)| = O\left(\frac{\sqrt{\ln n}}{pn}\right)$$

Proof:

$$X = b(k;n,p) - b(k-1;n-1,p) = \binom{n}{k} p^k q^{n-k} - \binom{n-1}{k-1} p^{k-1} q^{n-k}$$

$$= \binom{n}{k} p^k q^{n-k} \left(1 - \frac{k}{pn}\right) = b(k;n,p)\left(1 - \frac{k}{pn}\right)$$

Let $\delta = 1 - \frac{k}{pn}$, so $k = pn(1-\delta)$.

If $|\delta| \leq \sqrt{6 \ln n / pn}$, then, from lemma 3.1:

$$|X| = b(k;n,p)\,|\delta| \leq \Theta\left(\frac{1}{\sqrt{pn}}\right) \frac{\sqrt{6 \ln n}}{\sqrt{pn}} = O\left(\frac{\sqrt{\ln n}}{pn}\right)$$

Otherwise, $|\delta| > \sqrt{6 \ln n/pn}$ and we use Chernoff's bound [10] for the tails of binomial distribution:

$$|X| = b(k;n,p)\,|\delta| \le n \cdot b(k;n,p) = n \Pr(B(n,p) = pn(1-\delta))$$
$$\le n \Pr(|B(n,p) - pn| \ge pn\,|\delta|) \le 2ne^{-\delta^2 pn/3} < 2ne^{-2\ln n} = \frac{2}{n} \le \frac{1}{pn} = O\left(\frac{\sqrt{\ln n}}{pn}\right)$$

This gives us the first bound.

The second bound follows, because:

$$b(k;n,p) = pb(k-1;n-1,p) + qb(k;n-1,p)$$

Hence $b(k;n,p)$ is closer to $b(k;n-1,p)$ than to $b(k-1;n-1,p)$.

The third bound follows from the first two by triangle inequality. □

Lemma 3.3 *If $p = \omega(\ln n/n)$, $|k - pn| = o\left(\sqrt{pn/\ln n}\right)$, then:*

$$b(k;n,p) = \Theta\left(\frac{1}{\sqrt{pn}}\right)$$

Proof: This follows from lemma 3.1 and lemma 3.2 (third inequality) applied $|k - pn|$ times:

$$b(k;n,p) = b(\lfloor pn \rfloor;n,p) \pm |k - \lfloor pn \rfloor| \cdot O\left(\frac{\sqrt{\ln n}}{pn}\right)$$
$$= \Theta\left(\frac{1}{\sqrt{pn}}\right) \pm o\left(\sqrt{\frac{pn}{\ln n}}\right) \cdot O\left(\frac{\sqrt{\ln n}}{pn}\right) = \Theta\left(\frac{1}{\sqrt{pn}}\right)$$

□

Lemma 3.4 *If $n > 0$, $pn = \Omega(1)$:*

$$\Pr(B(n,p) = B'(n,p)) = O\left(\frac{1}{\sqrt{pn}}\right)$$

Proof: This follows directly from lemma 3.1:

$$\Pr(B(n,p) = B'(n,p)) = \sum_{i=0}^{n} b(i;n,p)^2 = O\left(\frac{1}{\sqrt{pn}}\right) \sum_{i=0}^{n} b(i;n,p) = O\left(\frac{1}{\sqrt{pn}}\right)$$

□

Now we can proceed to prove theorems about our algorithm.

Theorem 3.5 *Let a, b be two distinct vertices of the graph G with equal degree neighborhoods. Let $G' = G - \{a, b\}$ be the subgraph obtained by removing vertices a and b from G. Then the multiset of the G'-degrees of the vertices in G' connected to a is equal to the multiset of the G'-degrees of the vertices in G' connected to b.*

Proof: Since the degree neighborhoods of a and b are equal, the degrees of a and b are also equal (the lengths of neighborhoods are same).

Let A be the set of vertices connected to a in G, B be the set of vertices connected to b in G. A and B "generate" the same degree multisets.

Let $A' = A \cap G'$, $B' = B \cap G'$. If a and b are not connected, then $A = A'$, $B = B'$. If they are connected, then $A' = A - b$, $B' = B - a$. Since the degrees of a and b are equal, A' and B' in both cases generate the same G-degree multisets.

Let $C = A' \cap B'$, $A'' = A' - C$, $B'' = B' - C$. Then $A' = C \cup A''$, $B' = C \cup B''$. Therefore, A'' and B'' generate the same G-degree multisets. But all vertices in $A'' \cup B''$ are connected to exactly one of a, b. Therefore, the G'-degrees in $A'' \cup B''$ are one less than the G-degrees. Thus A'' and B'' generate the same G'-degree multisets. Hence also $A' = A'' \cup C$ and $B' = B'' \cup C$ generate the same G'-degree multisets. $\square$

The next theorem is about proving the existence of some number of vertices whose degrees lie in the range $[D_0, D_0 + R - 1]$ for appropriate D_0 and R.

Theorem 3.6 *If $p = \omega(\ln^3 n/n)$, $|D_0 - pn| = o\left(\sqrt{pn/\ln n}\right)$, $R = o\left(\sqrt{pn/\ln n}\right)$, $R = \omega\left(\sqrt{\ln n \ln(pn)}\right)$ then with high probability there exist in $G(n,p)$ at least $\left\lfloor \sqrt{\frac{n}{p} \ln n \ln(pn)} \right\rfloor$ vertices with degrees in the range $[D_0, D_0 + R - 1]$.*

Proof: Let X be a random variable denoting the number of such vertices. Let $X = X_1 + \ldots + X_n$, where each X_i is a 0-1 random variable equal to 1 if vertex number i has degree in the given range.

Since the distribution of the degree of a vertex is $B(n-1, p)$ and from lemma 3.3 we know that the whole range $[D_0, D_0 + R - 1]$ falls in the range of highest probability, we have:

$$\mathrm{E}[X_i] = \sum_{d=D_0}^{D_0+R-1} b(d; n-1, p) = R \cdot \Theta\left(\frac{1}{\sqrt{pn}}\right) = \omega\left(\frac{\sqrt{\ln n \ln(pn)}}{\sqrt{pn}}\right)$$

$$\mathrm{E}[X] = n\,\mathrm{E}[X_i] = \omega\left(\sqrt{\frac{n}{p} \ln n \ln(pn)}\right)$$

Now our goal is to prove that $X \geq \mathrm{E}[X]/2$ with high probability. To do that we will use a technique of proving concentration of random variables around the mean using a Doob's martingale [10, 11].

Let's build an edge-exposure martingale [10] for our graph. The martingale will represent the evolution of the expected value for X as we randomly decide for each edge whether or not to include it in the graph.

Number the vertices from 1 to n and number the possible edges in the order $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, $\{1,4\}$, $\{2,4\}$, $\{3,4\}$ ..., $\{n-1,n\}$. In other words, first we connect 2 to the smaller vertices (1), then connect 3 to the smaller vertices (1, 2), then we connect 4, and so on.

Define random variables $C_i = 1$ if the edge number i is chosen, 0 otherwise. Define the Doob's martingale [11] $Z_i = \mathrm{E}[X|C_1, \ldots, C_i]$. Clearly $Z_0 = \mathrm{E}[X]$, $Z_{\binom{n}{2}} = X$.

We want to use Azuma's Inequality [11] to prove a probabilistic lower bound on X. It states that if $|Z_i - Z_{i-1}| \le z_i$ then:

$$\Pr(X \le \mathrm{E}[X] - t) = \Pr\left(Z_{\binom{n}{2}} \le Z_0 - t\right) \le e^{-t^2/\left(2\sum z_i^2\right)} \tag{2}$$

We will use $t = \mathrm{E}[X]/2$. We need a good upper bound on $|Z_i - Z_{i-1}|$.

Let the edge number i be $\{a,b\}$, $a < b$.

$$\begin{aligned}
|Z_i - Z_{i-1}| &= \left|\mathrm{E}[X|C_1,\ldots,C_i] - \mathrm{E}[X|C_1,\ldots,C_{i-1}]\right| \\
&= \left|\sum_{v=1}^{n}\left(\mathrm{E}[X_v|C_1,\ldots,C_i] - \mathrm{E}[X_v|C_1,\ldots,C_{i-1}]\right)\right| \\
&\le \sum_{v=1}^{n}\left|\mathrm{E}[X_v|C_1,\ldots,C_i] - \mathrm{E}[X_v|C_1,\ldots,C_{i-1}]\right| \\
&= \left|\mathrm{E}[X_a|C_1,\ldots,C_i] - \mathrm{E}[X_a|C_1,\ldots,C_{i-1}]\right| \\
&\quad + \left|\mathrm{E}[X_b|C_1,\ldots,C_i] - \mathrm{E}[X_b|C_1,\ldots,C_{i-1}]\right|
\end{aligned}$$

since X_v is independent from C_i for $v \ne a, b$.

Let w be the number of chosen edges incident to b among $C_1, \ldots, C_{i-1}$. Let y be the number of remaining possible edges among $C_{i+1}, \ldots, C_{\binom{n}{2}}$ incident to b. Clearly $y \ge n - b$, because of the order in which the edges are taken (vertices bigger than b have not yet been connected).

Let $B = \left|\mathrm{E}[X_b|C_1,\ldots,C_i] - \mathrm{E}[X_b|C_1,\ldots,C_{i-1}]\right|$.

If $C_i = 0$ then:

$$\begin{aligned}
B &= \left|\mathrm{E}[X_b|C_1,\ldots,C_i] - \mathrm{E}[X_b|C_1,\ldots,C_{i-1}]\right| \\
&= \left|\sum_{d=D_0}^{D_0+R-1} b(d-w;y,p) - \sum_{d=D_0}^{D_0+R-1} b(d-w;y+1,p)\right| \\
&= \left|\sum_{d=D_0}^{D_0+R-1} b(d-w;y,p) - \sum_{d=D_0}^{D_0+R-1} \left(pb(d-w-1;y,p) + (1-p)b(d-w;y,p)\right)\right| \\
&= \left|\sum_{d=D_0}^{D_0+R-1} b(d-w;y,p) - p\sum_{d=D_0-1}^{D_0+R-2} b(d-w;y,p) - (1-p)\sum_{d=D_0}^{D_0+R-1} b(d-w;y,p)\right| \\
&= \left|pb(D_0-w-1;y,p) - pb(D_0-w-1+R;y,p)\right| \\
&\le b(D_0-w-1;y,p) + b(D_0-w-1+R;y,p)
\end{aligned}$$

Similarly if $C_i = 1$, then:

$$\begin{aligned}
B &= \left|\mathrm{E}\left[X_b|C_1,\ldots,C_i\right] - \mathrm{E}\left[X_b|C_1,\ldots,C_{i-1}\right]\right| \\
&= \left|\sum_{d=D_0}^{D_0+R-1} b(d-w-1;y,p) - \sum_{d=D_0}^{D_0+R-1} b(d-w;y+1,p)\right| \\
&= \left|\sum_{d=D_0}^{D_0+R-1} b(d-w-1;y,p) - \sum_{d=D_0}^{D_0+R-1} \big(pb(d-w-1;y,p) - (1-p)b(d-w;y,p)\big)\right| \\
&= \left|\sum_{d=D_0-1}^{D_0+R-2} b(d-w;y,p) - p\sum_{d=D_0-1}^{D_0+R-2} b(d-w;y,p) - (1-p)\sum_{d=D_0}^{D_0+R-1} b(d-w;y,p)\right| \\
&= \left|(1-p)b(D_0-w-1;y,p) - (1-p)b(D_0-w-1+R;y,p)\right| \\
&\leq b(D_0-w-1;y,p) + b(D_0-w-1+R;y,p)
\end{aligned}$$

In both cases:

$$B \leq b(D_0-w-1;y,p) + b(D_0-w-1+R;y,p)$$

If $y+1 \geq 1/p$ then using lemma 3.1 we have:

$$B \leq \frac{C}{\sqrt{p(y+1)}}$$

for a large enough constant C.

If $y+1 < 1/p$ then we will use the obvious bound $B \leq 2$.

Therefore if we make sure $C \geq 2$ we have:

$$B \leq C\min\left(1, \frac{1}{\sqrt{p(y+1)}}\right) \leq C\min\left(1, \frac{1}{\sqrt{p(n-b+1)}}\right)$$

Above reasoning can be repeated for $A = \left|\mathrm{E}\left[X_a|C_1,\ldots,C_i\right] - \mathrm{E}\left[X_a|C_1,\ldots,C_{i-1}\right]\right|$. In this case we will have y^* remaining edges not yet connected to a with $y^* \geq n-b$, and taking large enough C':

$$A \leq C'\min\left(1, \frac{1}{\sqrt{p(y^*+1)}}\right) \leq C'\min\left(1, \frac{1}{\sqrt{p(n-b+1)}}\right)$$

Therefore $|Z_i - Z_{i-1}| \leq z_i$ where:

$$z_i = (C + C') \min\left(1, \frac{1}{\sqrt{p(n-b+1)}}\right)$$

$$\sum z_i^2 = (C+C')^2 \sum_{b=2}^{n} \sum_{a=1}^{b-1}$$

$$\leq (C+C')^2 n \sum_{b=2}^{n} \min\left(1, \frac{1}{\sqrt{p(n-b+1)}}\right)^2 \leq (C+C')^2 n \sum_{u=1}^{n} \min\left(1, \frac{1}{pu}\right)$$

$$\leq (C+C')^2 \frac{n}{p} \sum_{u=1}^{n} \min(p, \frac{1}{u}) = (C+C')^2 \frac{n}{p} \left(\sum_{u=1}^{\lfloor 1/p \rfloor} p + \sum_{u=\lfloor 1/p \rfloor + 1}^{n} \frac{1}{u}\right)$$

$$= O\left(\frac{n}{p}(1 + \ln n - \ln(1/p))\right) = O\left(\frac{n}{p} \ln(pn)\right)$$

Using Azuma's inequality (Equation 2), we have:

$$\Pr(X \leq \mathrm{E}[X]/2) \leq \exp\left(-\frac{(\mathrm{E}[X]/2)^2}{2\sum z_i^2}\right)$$

$$= \exp\left(-\frac{\omega\left(\sqrt{n \ln n \ln(pn)/p}\right)^2}{O(n \ln(pn)/p)}\right) = \exp\left(-\omega(\ln n)\right)$$

Hence with high probability $X > \mathrm{E}[X]/2 > \left\lfloor \sqrt{\frac{n}{p} \ln n \ln(pn)} \right\rfloor$. □

In the corollary below we show a lower bound on how many disjoint degree ranges we can take so that each range has some vertices in it with high probability.

Corollary 3.7 *If* $p = \omega(\ln^3 n/n)$ *and* $x = o\left(\sqrt{pn/(\ln^2 n \ln(pn))}\right)$ *then we can find* x *nonoverlapping ranges of degrees of length* $R = \omega\left(\sqrt{\ln n \ln(pn)}\right)$ *such that in* $G(n,p)$ *there will be at least* $K = \left\lfloor \sqrt{n \ln n \ln(pn)/p} \right\rfloor$ *vertices in each range with high probability.*

Proof: Since:

$$\frac{\sqrt{pn/\ln n}}{x} = \omega(\sqrt{\ln n \ln(pn)})$$

we can find R such that:

$$R = \omega\left(\sqrt{\ln n \ln(pn)}\right)$$

$$Rx = o\left(\sqrt{pn/\ln n}\right)$$

This follows from the general theorem that if $f(n) = o(g(n))$ then we can find $h(n)$ such that $h(n) = o(g(n))$ and $h(n) = \omega(f(n))$. For example, $h = \sqrt{fg}$ will do.

Now just find x separate ranges of length R around pn with the distance from pn of the order $o\left(\sqrt{pn/\ln n}\right)$ and apply the theorem. $\square$

Now we can use the corollary to estimate the failure probability of the algorithm.

Theorem 3.8 *If $p \leq 1/2$, $p = \omega\left(\ln^4 n/n \ln\ln n\right)$ then the probability that the algorithm fails is small, i.e., $O(n^{-\alpha})$, for every constant $\alpha > 0$.*

Proof: Since:

$$\frac{pn}{\ln(pn)} = \omega\left(\frac{\ln^4 n/\ln\ln n}{\ln\ln n}\right) = \omega\left(\frac{\ln^4 n}{\ln^2 \ln n}\right)$$
$$\frac{\ln^2 n}{\ln^2 \ln n} = o\left(\frac{pn}{\ln^2 n \ln(pn)}\right)$$
$$\frac{\ln n}{\ln\ln n} = o\left(\sqrt{\frac{pn}{\ln^2 n \ln(pn)}}\right)$$

we can find x such that:

$$x = o\left(\sqrt{\frac{pn}{\ln^2 n \ln(pn)}}\right)$$
$$x = \omega\left(\frac{\ln n}{\ln\ln n}\right)$$

Take any two vertices a, b in the graph G. Let $G' = G - \{a, b\}$. G' is a random $G(n-2, p)$ graph, so according to the corollary above, we can find x disjoint ranges of degrees such that with high probability there exist in G' at least $K = \left\lfloor \sqrt{n' \ln n' \ln(pn')/p} \right\rfloor$ (where $n' = n-2$) vertices with degrees falling in each range.

If a and b are to have the same degree neighborhoods, then from theorem 3.5, for every range both a and b must be connected to the same number of vertices in that range. Since $Kp = \omega(1)$, from lemma 3.4, the probability of that happening for a given group is at most:

$$O\left(\frac{1}{\sqrt{Kp}}\right) = O\left((pn \ln n \ln(pn))^{-1/4}\right)$$
$$= O\left((pn)^{-1/4}\right) = \exp(-\Omega(\ln(pn))) = \exp(-\Omega(\ln\ln n))$$

Connections to each group of vertices are independent, therefore the probability of a and b having the same degree neighborhoods is bounded by:

$$(\exp(-\Omega(\ln\ln n)))^x = \exp(-\Omega(x \ln\ln n)) = \exp(-\omega(\ln n))$$

There are fewer than n^2 such pairs (a, b) and each pair has same degree neighborhood with small probability. By the union bound ([10, Lemma 1.2]), the probability that any pair

of vertices having the same degree neighborhood is at most n^2 times the above probability which is (still) bounded by $\exp(-\omega(\ln n)) = O(n^{-\alpha})$ (for every constant $\alpha > 0$). The algorithm will fail only if the degree neighborhoods are not distinct for each vertex (Step 4). Hence the algorithm will succeed with high probability. □

The theorems 2.1, 2.2 and 3.8 show that our algorithm is correct, runs in linear time and succeeds with high probability for $p = \omega(\ln^4 n/n \ln \ln n)$.

References

[1] L. Babai, P. Erdös, and S. M. Selkow. Random Graph Isomorphism, *SIAM J. Computing*, **9**(3), 1980, 628–635.

[2] L. Babai and L. Kučera, Canonical Labelling of Graphs in Linear Average Time, *Proc. of 20th Annual IEEE Symp. on Foundations of Computational Science*, Puerto Rico, 1979, 39–46.

[3] L. Babai and E. Luks. Canonical labeling of graphs, *Proc. of 15th ACM Symp. on Theory of Computing*, 1983, 171–183.

[4] B. Bollobás. *Random Graphs*, Cambridge University Press, 2001.

[5] B. Bollobás. Distinguishing vertices of random graphs, *Annals Discrete Math*, **13**, 33-50.

[6] A. Frieze and C. McDiarmid. Algorithmic theory of random graphs, *Random Structures and Algorithms*, **10**, 1997, 5-42.

[7] R. Karp. Probabilistic analysis of a Canonical Numbering Algorithm for Graphs, *Proc. Symposia in Pure Math.*, **34**, American Mathematical Society, Providence, RI, 1979, 365-378.

[8] D. Knuth. *The Art of Programming: Sorting and Searching*, Addison-Wesley, 1998.

[9] R. Lipton. The Beacon Set Approach to Graph Isomorphism, Yale University, preprint no. 135, (1978).

[10] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press, 2005.

[11] R. Motwani and P. Raghavan. *Randomized Algorithms*, Cambridge University Press, 1995.