Canonization of a Random Graph by Two Matrix-Vector Multiplications

Oleg Verbitsky
Institut für Informatik, Humboldt-Universität zu Berlin, Germany

Maksim Zhukovskii
Department of Computer Science, University of Sheffield, UK

Abstract

We show that a canonical labeling of a random n-vertex graph can be obtained by assigning to each vertex x the triple $(w_1(x), w_2(x), w_3(x))$, where $w_k(x)$ is the number of walks of length k starting from x. This takes time $\mathcal{O}(n^2)$, where n^2 is the input size, by using just two matrix-vector multiplications. The linear-time canonization of a random graph is the classical result of Babai, Erdős, and Selkow. For this purpose they use the well-known combinatorial color refinement procedure, and we make a comparative analysis of the two algorithmic approaches.

2012 ACM Subject Classification Mathematics of computing → Random graphs; Mathematics of computing → Graph algorithms

Keywords and phrases Graph Isomorphism, canonical labeling, random graphs, walk matrix, color refinement, linear time

Digital Object Identifier 10.4230/LIPIcs.ESA.2023.100

Funding *Oleg Verbitsky*: Supported by DFG grant KO 1053/8–2. On leave from the IAPMM, Lviv, Ukraine.

1 Introduction

A *walk* in a graph $G = (V, E)$ is a sequence of vertices $x_0x_1 \dots x_k$ such that $(x_i, x_{i+1}) \in E$ for every $0 \le i < k$. We say that $x_0x_1 \dots x_k$ is a walk of *length* k *from* x_0 *to* x_k. For a vertex $x \in V$, let $w_k^G(x)$ denote the total number of walks of length k in G starting from x. Furthermore, we define $\mathbf{w}_k^G(x) = (w_1^G(x), \dots, w_k^G(x))$.

The Erdős-Rényi random graph $G(n, p)$ is a graph on the vertex set $[n] = \{1, \dots, n\}$ where each pair of distinct vertices x and y is adjacent with probability p independently of the other pairs. In particular, $G(n, 1/2)$ is a random graph chosen equiprobably from among all graphs on $[n]$.

▶ **Theorem 1.** *Let $G = G(n, 1/2)$. Then*

$$\mathbf{w}_3^G(x) \neq \mathbf{w}_3^G(y) \text{ for all } x \neq y$$

with probability at least $1 - O(\sqrt[4]{\ln n / n})$.

If α is an isomorphism from a graph G to a graph H, then clearly $\mathbf{w}_k^G(x) = \mathbf{w}_k^H(\alpha(x))$. Theorem 1, therefore, shows that the map $x \mapsto \mathbf{w}_3^G(x)$ is a canonical labeling of G for almost all n-vertex graphs G. This labeling is easy to compute. Indeed, if A is the adjacency matrix of G and $\mathbb{1}$ is the all-ones vector-column of length n, then

$$(w_k^G(1), \dots, w_k^G(n))^\top = A^k \mathbb{1}.$$

After noting that $w_1^G(x) = d(x)$, where $d(x)$ denotes the degree of a vertex x, this yields the following simple canonical labeling algorithm.

Algorithm A. Canonical labeling of a random graph.

INPUT: a graph G on $[n]$ with adjacency matrix A.

1. Form a vector $D_1 = (d(1), \dots, d(n))^\top$.
2. Compute $D_2 = AD_1$ and $D_3 = AD_2$.
3. Let W be the matrix formed by the three columns D_1, D_2, D_3 and let $W_1, \dots, W_n$ be the rows of W.
4. If there are identical rows $W_x = W_y$ for some $x \neq y$, then give up. Otherwise,
5. to each vertex x, assign the label W_x.

▶ **Corollary 2.** *Algorithm A with high probability canonizes a random n-vertex graph, taking time $\mathcal{O}(n^2)$ on every input.*

The notation $\mathcal{O}(\cdot)$ in the time bound means a linear function up to the logarithmic factor $\log n \log\log n$ corresponding to the complexity of integer multiplication [7], that is, $\mathcal{O}(n^2) = O(n^2 \log n \log\log n)$. If the model of computation assumes that multiplication of two integers takes a constant time, then we just set $\mathcal{O}(n^2) = O(n^2)$. This time bound stems from the fact that the two matrix-vector multiplications in Step 2 are the most expensive operations performed by the algorithm. Note that this bound is essentially linear because a random graph is with high probability dense, having $(1/4 + o(1))n^2$ edges.

The linear-time canonization of almost all graphs is a classical result of Babai, Erdős, and Selkow [2], which was a basis for settling the average-case complexity of graph isomorphism in [3]. While our algorithm is based solely on basic linear-algebraic primitives, the method used in [2, 3] is purely combinatorial. Before comparing the two approaches, we put Theorem 1 in the context of the earlier work on walk counts and their applications to isomorphism testing.

Of course, Algorithm A can be enhanced by taking into account also longer walks, that is, by involving also other vector-columns $A^k\mathbb{1}$ for $k > 3$. Note that there is no gain in considering these vectors for $k \geq n$. Indeed, if $A^k\mathbb{1}$ is a linear combination of the vectors $\mathbb{1}, A\mathbb{1}, A^2\mathbb{1}, \dots, A^{k-1}\mathbb{1}$, the same is obviously true also for $A^{k+1}\mathbb{1}$ (cf. [11, Lemma 1] and see also [6] for a more detailed linear-algebraic analysis). Therefore, it suffices to start our matrix W from the column $\mathbb{1}$ and add a subsequent column $A^k\mathbb{1}$ as long as this increases the rank of W, which is possible only up to $k = n - 1$. The $n \times n$ matrix $W = W^G$ formed by the columns $\mathbb{1}, A\mathbb{1}, \dots, A^{n-1}\mathbb{1}$ is called the *walk matrix* of the graph G (*WM* for brevity). The entries of $W^G = (w_{x,k})_{1\leq x\leq n,\, 0\leq k<n}$ are nothing else as the walk counts $w_{x,k} = w_k^G(x)$. Note that $w_0^G(x) = 1$ as there is a single walk of length 0 from x.

We say that a graph G is *WM-discrete* if the rows of the walk matrix W^G are pairwise different, i.e., $W_x^G \neq W_y^G$ for all $x \neq y$. For a such G, the walk matrix yields a canonical labeling where each vertex x is assigned the vector $W_x^G = (w_0^G(x), w_1^G(x), \dots, w_{n-1}^G(x))$. Note that if W^G has identical rows, then this matrix is singular; cf. [5, Section 7]. O'Rourke and Touri [10] prove that the walk matrix of a random graph is non-singular with high probability. This implies that a random graph is WM-discrete with high probability and, hence, almost all graphs are canonizable by computing the $n \times n$ walk matrix similarly to Algorithm A. Note that this takes time $O(n^3)$, which is outperformed by our Corollary 2 due to using the truncated variant of WM of size $n \times 4$.

Remarkably, non-singular walk matrices can be used to test isomorphism of two given graphs directly rather than by computing their canonical forms. If graphs G and H are isomorphic, then their walk matrices W^G and W^H can be obtained from one another by

rearranging the rows. If the last condition is satisfied, we write $W^G \approx W^H$. This relation between matrices is efficiently checkable just by sorting the rows in the lexicographic order. We say that a graph G is *WM-identifiable* if, conversely, for all H we have $G \cong H$ whenever $W^G \approx W^H$. Liu and Siemons [8] prove that if the walk matrix of a graph is non-singular, then it uniquely determines the adjacency matrix. This implies by [10] that a random graph is WM-identifiable with high probability.

Note that, by a simple counting argument, almost all n-vertex graphs *cannot* be identified by the shorter version of the walk matrix of size $n \times k$ as long as $k = o(\sqrt{n/\log n})$. In particular, Theorem 1 cannot be extended to the identifiability concept.

The combinatorial approach of Babai, Erdős, and Selkow [2] is based on the *color refinement* procedure (*CR* for brevity) dating back to the sixties (e.g., [9]). CR begins with a uniform coloring of all vertices in an input graph and iteratively refines a current coloring according to the following principle: If two vertices are equally colored but have distinct color frequencies in their neighborhoods, then they get distinct colors in the next refinement step. The refinement steps are executed as long as the refinement is proper. As soon as the color classes stay the same, CR terminates and outputs the current coloring (a detailed description of the algorithm is given in Section 3.1). CR *distinguishes* graphs G and H if their color palettes are distinct. A graph G is called *CR-identifiable* if it is distinguishable by CR from every non-isomorphic H. CR can also be used for computing a canonical labeling of a single input graph. We say that a graph G is *CR-discrete* if CR assigns a unique color to each vertex of G. It is easy to prove that every CR-discrete graph is CR-identifiable. We do not know whether or not this is true also for the corresponding WM concepts.

Powers and Sulaiman [11] discuss examples when the CR-partition and the WM-partition are different, that is, CR and the WM-based vertex-classification algorithm give different results. In particular, [11, Fig. 3] shows a graph which is, in our terminology, CR-discrete but not WM-discrete. We give a finer information about the relationship between the two algorithmic approaches.

▶ **Theorem 3.**

1. *Every WM-discrete graph is also CR-discrete.*
2. *Every WM-identifiable graph is also CR-identifiable.*
3. *There is a graph that is*
   - **(a)** *CR-discrete (hence also CR-identifiable) and*
   - **(b)** *neither WM-discrete*
   - **(c)** *nor WM-identifiable.*

Theorem 3 shows that the WM approach is superseded by the CR algorithm with regard to canonization of a single input graph and testing isomorphism of two input graphs. Moreover, CR is sometimes more successful with respect to both algorithmic problems. Thus, WM can be regarded as a weaker algorithmic tool for canonical labeling and isomorphism testing, which is not so surprising as this approach is actually based on a single basic linear-algebraic primitive, namely matrix-vector multiplication. In this sense, Algorithm A is arguably simpler than the classical CR-based canonization of a random graph as it demonstrates that a random graph can be canonized in an essentially linear time even with less powerful computational means.

Theorems 1 and 3 are proved in Sections 2 and 3 respectively.

2 Canonization of a random graph

2.1 Probability preliminaries

Let X be a binomial random variable with parameters n and p, that is, $X = \sum_{i=1}^{n} X_i$ where X_i's are mutually independent and, for each i, we have $X_i = 1$ with probability 1 and $X_i = 0$ with probability $1 - p$. We use the notation $X \sim \mathrm{Bin}(n, p)$ when X has this distribution. As well known, X is well-concentrated around its expectation np.

▶ **Lemma 4** (Chernoff's bound; see, e.g., [1, Corollary A.1.7]). *If $X \sim \mathrm{Bin}(n, p)$, then*

$$\mathsf{P}[|X - np| > t] \le 2e^{-2t^2/n}$$

for every $t \ge 0$.

▶ **Lemma 5.** *If X and Y are independent random variables, each having the probability distribution $\mathrm{Bin}(n, 1/2)$, then $\mathsf{P}[X = Y] < 1/\sqrt{\pi n}$.*

Proof. Using the well-known estimate

$$\binom{2n}{n} < \frac{2^{2n}}{\sqrt{\pi n}}, \tag{1}$$

we obtain

$$\mathsf{P}[X = Y] = \sum_{k=0}^{n} \left(\binom{n}{k} 2^{-n} \right)^2 = 2^{-2n} \sum_{k=0}^{n} \binom{n}{k}^2 = 2^{-2n} \binom{2n}{n} < \frac{1}{\sqrt{\pi n}},$$

where the last equality is a special case of Vandermonde's convolution. ◀

2.2 Proof of Theorem 1

For a vertex $i \in [n]$, recall that $\mathbf{w}_3(i) = (w_1^G(i), w_2^G(i), w_3^G(i))$. By the union bound,

$$\mathsf{P}[\mathbf{w}_3(i) = \mathbf{w}_3(j) \text{ for some } i, j] \le \sum_{i,j} \mathsf{P}[\mathbf{w}_3(i) = \mathbf{w}_3(j)] = \binom{n}{2} \mathsf{P}[\mathbf{w}_3(1) = \mathbf{w}_3(2)].$$

Therefore, it suffices to prove that

$$\mathsf{P}[\mathbf{w}_3(1) = \mathbf{w}_3(2)] = O(n^{-9/4} \ln^{1/4} n). \tag{2}$$

Let $N_H(v)$ denote the neighborhood of a vertex v in a graph H. Given two sets $U_1 \subset [n] \setminus \{1\}$ and $U_2 \subset [n] \setminus \{2\}$, let $G' = G'(U_1, U_2)$ be the random graph G subject to the conditions $N_G(1) = U_1$ and $N_G(2) = U_2$. In other terms, G' is a random graph on $[n]$ chosen equiprobably among all graphs satisfying these conditions. Let $w'_k(i) = w_k^{G'}(i)$ denote the number of walks of length k emanating from i in G' (the dependence of $w'_k(i)$ on the pair U_1, U_2 will be dropped for the sake of notational convenience). Define

$$p(U_1, U_2) = \mathsf{P}\left[\sum_{i \in U_1} w'_1(i) = \sum_{i \in U_2} w'_1(i) \text{ and } \sum_{i \in U_1} w'_2(i) = \sum_{i \in U_2} w'_2(i) \right].$$

We have

$$\begin{aligned}&\mathsf{P}[\mathbf{w}_3(1)=\mathbf{w}_3(2)]\\&=\sum_{U_1,U_2\,:\,|U_1|=|U_2|}\mathsf{P}[\mathbf{w}_3(1)=\mathbf{w}_3(2)\mid N_G(1)=U_1,\ N_G(2)=U_2]\\&\qquad\times\mathsf{P}[N_G(1)=U_1,\ N_G(2)=U_2]\\&\qquad=\sum_{U_1,U_2\,:\,|U_1|=|U_2|}p(U_1,U_2)\times\mathsf{P}[N_G(1)=U_1,\ N_G(2)=U_2].\quad(3)\end{aligned}$$

Note first that

$$\begin{aligned}\sum_{|U_1|=|U_2|}\mathsf{P}\left[N_G(1)=U_1,\ N_G(2)=U_2\right]&=\mathsf{P}\left[|N_G(1)|=|N_G(2)|\right]\\&=\mathsf{P}\left[|N_G(1)\setminus\{2\}|=|N_G(2)\setminus\{1\}|\right]=O(n^{-1/2})\end{aligned}$$

by Lemma 5 because $|N_G(1)\setminus\{2\}|\sim\mathrm{Bin}(n-2,1/2)$ and $|N_G(2)\setminus\{1\}|\sim\mathrm{Bin}(n-2,1/2)$ are independent binomial random variables. This allows us to derive (2) from (3) if we prove that

$$p(U_1,U_2)=O(n^{-7/4}\ln^{1/4}n)\tag{4}$$

for the neighborhood sets U_1 and U_2.

In fact, we do not need to prove (4) for all pairs U_1,U_2 because the contribution of some of them in (3) is negligible. Indeed, set $\varepsilon(n)=n^{-1/4}$. Note that $|N_G(j)|\sim\mathrm{Bin}(n-1,1/2)$ for $j=1,2$ and $|(N_G(1)\cap N_G(2))\setminus\{1,2\}|\sim\mathrm{Bin}(n-2,1/4)$. By the Chernoff bound (see Lemma 4), we have $(1/2-\varepsilon(n))n\le|N_G(j)|\le(1/2+\varepsilon(n))n$ for $j=1,2$ and $(1/4-\varepsilon(n))n\le|N_G(1)\cap N_G(2)|\le(1/4+\varepsilon(n))n$ with probability $1-e^{-\Omega(\sqrt{n})}$. Call a pair U_1,U_2 *standard* if $|U_j|$ for $j=1,2$ and $|U_1\cap U_2|$ are in the same ranges. Thus, all non-standard pairs make a negligible contribution in (3), and we only have to prove (4) for each standard pair U_1,U_2.

For a graph H and a subset $U\subset V(H)$, let $E_H(U)$ denote the set of edges of H with at least one vertex in U. Given two sets of edges $\mathcal{E}_1$ and $\mathcal{E}_2$ incident to the vertices in $U_1\setminus\{2\}$ and $U_2\setminus\{1\}$ respectively, let $G''=G''(U_1,U_2,\mathcal{E}_1,\mathcal{E}_2)$ be the random graph G' subject to the conditions $E_{G'}(U_1\setminus\{2\})=\mathcal{E}_1$ and $E_{G'}(U_2\setminus\{1\})=\mathcal{E}_2$. Let $w''_k(i)=w_k^{G''}(i)$ denote the number of walks of length k emanating from i in G'' (the dependence of $w''_k(i)$ on $U_1,U_2,\mathcal{E}_1,\mathcal{E}_2$ is dropped for notational simplicity). Using this notation, we can write

$$\begin{aligned}p(U_1,U_2)=\sum_{\mathcal{E}_1,\mathcal{E}_2:\,\sum_{U_1}w'_1(i)=\sum_{U_2}w'_1(i)}&\mathsf{P}\left[E_{G'}(U_1\setminus\{2\})=\mathcal{E}_1,\ E_{G'}(U_2\setminus\{1\})=\mathcal{E}_2\right]\\&\times\mathsf{P}\left[\sum_{i\in U_1}w''_2(i)=\sum_{i\in U_2}w''_2(i)\right].\quad(5)\end{aligned}$$

We first show that

$$\sum_{\mathcal{E}_1,\mathcal{E}_2:\,\sum_{U_1}w'_1(i)=\sum_{U_2}w'_1(i)}\mathsf{P}\left[E_{G'}(U_1\setminus\{2\})=\mathcal{E}_1,\ E_{G'}(U_2\setminus\{1\})=\mathcal{E}_2\right]=O(1/n).\tag{6}$$

Note that the sum in the left hand side of (6) is equal to the probability that $\sum_{i\in U_1}w'_1(i)=\sum_{i\in U_2}w'_1(i)$. This equality is equivalent to $\sum_{i\in U_1\setminus(U_2\cup\{2\})}w'_1(i)=\sum_{i\in U_2\setminus(U_1\cup\{1\})}w'_1(i)$, which in its turn is true if and only if $U_1\setminus(U_2\cup\{2\})$ and $U_2\setminus(U_1\cup\{1\})$ send the same number of edges to $[n]\setminus[(U_1\cup U_2\cup\{1,2\})\setminus(U_1\cap U_2)]$. Since the pair U_1,U_2 is standard, these numbers are independent binomial random variables with $\Theta(n^2)$ trials. Equality (6) now follows by Lemma 5.

We now can derive Equality (4) from Equality (6) by proving that

$$\mathsf{P}\left[\sum_{i\in U_1} w_2''(i) = \sum_{i\in U_2} w_2''(i)\right] = O(n^{-3/4}\ln^{1/4} n) \tag{7}$$

for each potential pair $\mathcal{E}_1, \mathcal{E}_2$. Again, it is enough to do this only for most probable pairs whose contribution in (5) is overwhelming. Specifically, let $w_2'(u,v)$ denote the number of all paths of length 2 between u and v in G' and define

$$\Delta(i,j) = (w_2'(i,1) + w_2'(j,1)) - (w_2'(i,2) + w_2'(j,2)).$$

Note that the numbers $w_2'(i,1), w_2'(j,1), w_2'(i,2), w_2'(j,2)$ and, hence, the numbers $\Delta(i,j)$ are completely determined by specifying $E_{G'}(U_1 \setminus \{2\}) = \mathcal{E}_1$ and $E_{G'}(U_2 \setminus \{1\}) = \mathcal{E}_2$. We call a pair $\mathcal{E}_1, \mathcal{E}_2$ *standard* if $\Delta(i,j)$ takes on $O(\sqrt{n\ln n})$ different values for $i \neq j$ from $[n] \setminus (U_1 \cup U_2 \cup \{1,2\})$. The following fact shows that it is enough if we prove (7) for each standard pair $\mathcal{E}_1, \mathcal{E}_2$.

▷ Claim 6. If a pair U_1, U_2 is standard, then

$$|\{\Delta(i,j) \,:\, i,j \in [n] \setminus (U_1 \cup U_2 \cup \{1,2\}),\, i \neq j\}| = O(\sqrt{n\ln n})$$

with probability $1 - O(n^{-6})$.

Proof. Let $u_1 = |U_1|$, $u_2 = |U_2|$, and $u = |U_1 \cap U_2|$. Note that

$$\Delta(i,j) = |N_{G'}(i) \cap (U_1 \setminus U_2)| + |N_{G'}(j) \cap (U_1 \setminus U_2)| - |N_{G'}(i) \cap (U_2 \setminus U_1)| - |N_{G'}(j) \cap (U_2 \setminus U_1)|.$$

The four terms in the right hand side are independent random variables $\mathrm{Bin}(u_1 - u, 1/2)$, $\mathrm{Bin}(u_1 - u, 1/2)$, $\mathrm{Bin}(u_2 - u, 1/2)$, $\mathrm{Bin}(u_2 - u, 1/2)$ respectively. Since $N - \mathrm{Bin}(N,p) \sim \mathrm{Bin}(N, 1-p)$, we conclude that $\Delta(i,j) \sim 2u - 2u_2 + \mathrm{Bin}(2u_1 + 2u_2 - 4u, 1/2)$. The Chernoff bound (see Lemma 4) implies that, for each pair i,j, the inequalities

$$2u - 2u_2 + (u_1 + u_2 - 2u)\left(1 - \frac{\sqrt{2\ln n}}{\sqrt{u_1 + u_2 - 2u}}\right) \leq \Delta(i,j) \leq 2u - 2u_2 + (u_1 + u_2 - 2u)\left(1 + \frac{\sqrt{2\ln n}}{\sqrt{u_1 + u_2 - 2u}}\right)$$

are violated with probability at most $O(n^{-8})$. By the union bound, the probability that not all values $\Delta(i,j)$ fall in an integer interval of length at most

$$2\sqrt{2\ln n(u_1 + u_2 - 2u)} = O(\sqrt{n\ln n})$$

is bounded by $O(n^{-6})$. ◁

It remains to prove (7) for a fixed standard pair $\mathcal{E}_1, \mathcal{E}_2$. Note that all walks of length 3 starting from 1 and 2 and having at least 2 vertices inside $U_1 \cup U_2 \cup \{1,2\}$ are determined by $U_1, U_2, \mathcal{E}_1, \mathcal{E}_2$. Let $\gamma_j = \gamma_j(U_1, U_2, \mathcal{E}_1, \mathcal{E}_2)$ denote the number of such walks starting at j for $j = 1, 2$. Let $e''_{i,j}$ be the indicator random variable of the presence of the edge $\{i,j\}$ in G''. The equality $\sum_{i\in U_1} w_2''(i) = \sum_{i\in U_2} w_2''(i)$ can be rewritten as

$$\gamma_1 + \sum_{i,j\notin U_1\cup U_2\cup\{1,2\}} e''_{ij}(w_2'(i,1) + w_2'(j,1)) = \gamma_2 + \sum_{i,j\notin U_1\cup U_2\cup\{1,2\}} e''_{ij}(w_2'(i,2) + w_2'(j,2)), \tag{8}$$

where the sums count the walks of length 3 from 1 and 2 whose last two vertices are outside $U_1 \cup U_2 \cup \{1,2\}$. Since $\mathcal{E}_1, \mathcal{E}_2$ is a standard pair, there exists an integer $x \neq 0$ such that $\Delta(i,j) = x$ for $\Omega(n^{3/2}/\sqrt{\ln n})$ pairs i,j. Let S_x be the set of all such pairs. Let $G^* = G^*(U_1, U_2, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}^*)$ be obtained from G'' by exposing all edges except those between i,j in S_x, where $\mathcal{E}^*$ is the set of exposed edges. Equality (8) is fulfilled if and only if

$$\sum_{\{i,j\} \in S_x} e''_{ij} x = \gamma(U_1, U_2, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}^*) \tag{9}$$

for some integer $\gamma(U_1, U_2, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}^*)$ which is completely determined by $U_1, U_2, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}^*$. It remains to note that the binomial random variable $\sum_{\{i,j\} \in S_x} e''_{ij} \sim \mathrm{Bin}(|S_x|, 1/2)$ takes on any fixed value with probability at most $\binom{|S_x|}{\lfloor |S_x|/2 \rfloor}/2^{|S_x|} = O(|S_x|^{-1/2}) = O(n^{-3/4} \ln^{1/4} n)$, where the first equality is due to (1). This completes the proof of Equality (7) and of the whole theorem.

▶ Remark 7. The probability bound in Theorem 1 cannot be significantly improved because $\mathsf{P}[\mathbf{w}_3^G(1) = \mathbf{w}_3^G(2)] = n^{-\Omega(1)}$. To see this, note first that $\mathsf{P}[w_1^G(1) = w_1^G(2)] = \Theta(n^{-1/2})$ (see the proof of Lemma 5). Assuming that a standard pair U_1, U_2 with $|U_1| = |U_2|$ is fixed, we can similarly show that $p(U_1, U_2) = \Theta(n^{-1})$, which implies that $\mathsf{P}[\big(w_1^G(1), w_2^G(1)\big) = \big(w_1^G(2), w_2^G(2)\big)] = \Theta(n^{-3/2})$. Showing a polynomial lower bound for $\mathsf{P}[\big(w_1^G(1), w_2^G(1), w_3^G(1)\big) = \big(w_1^G(2), w_2^G(2), w_3^G(2)\big)]$ is a slightly more delicate issue. Following the same proof strategy as for the upper bound, we have to ensure that the equation (9) has at least one integer solution $\sum_{\{i,j\} \in S_x} e''_{ij}$. We can do this because we have enough freedom in adjusting the right hand side of (9) by choosing an appropriate value of $\gamma_1 - \gamma_2$. Indeed, first of all, $|x|$ does not exceed $2n$ with probability 1. Second, we have an interval of length at least $100n$ for the values of $\gamma_1 - \gamma_2$ that are reachable with probability $\Omega(n^{-1})$. As easily seen, this is enough for obtaining a desired lower bound.

We leave as an open question whether Theorem 1 can be improved by excluding paths of length 3. Our conjecture is that this is impossible, that is, Theorem 1 is optimal in this respect, but proving this poses some technical challenges.

3 Comparing WM and CR

3.1 Color refinement

We begin with a formal description of the *color refinement* algorithm (*CR* for short). CR operates on vertex-colored graphs but applies also to uncolored graph by assuming that their vertices are colored uniformly. An input to the algorithm consists either of a single graph or a pair of graphs. Consider the former case first. For an input graph G with initial coloring C_0, CR iteratively computes new colorings

$$C_i(x) = \Big(C_{i-1}(x), \{\!\{C_{i-1}(y)\}\!\}_{y \in N(x)}\Big), \tag{10}$$

where $\{\!\{\}\!\}$ denotes a multiset and $N(x)$ is the neighborhood of a vertex x. Denote the partition of $V(G)$ into the color classes of C_i by $\mathcal{P}_i$. Note that each subsequent partition $\mathcal{P}_{i+1}$ is either finer than or equal to $\mathcal{P}_i$. If $\mathcal{P}_{i+1} = \mathcal{P}_i$, then $\mathcal{P}_j = \mathcal{P}_i$ for all $j \geq i$. Suppose that the color partition stabilizes in the t-th round, that is, t is the minimum number such that $\mathcal{P}_t = \mathcal{P}_{t-1}$. CR terminates at this point and outputs the coloring $C = C_t$. Note that if the colors are computed exactly as defined by (10), they will require exponentially long color names. To prevent this, the algorithm renames the colors after each refinement step, using the same set of no more than n color names.

We say that a graph G is *CR-discrete* if $C(x) \neq C(x')$ for all $x \neq x'$.

If an input consists of two graphs G and H, then it is convenient to assume that their vertex sets $V(G)$ and $V(H)$ are disjoint. The vertex colorings of G and H define an initial coloring C_0 of the union $V(G) \cup V(H)$, which is iteratively refined according to (10). The color partition $\mathcal{P}_i$ is defined exactly as above but now on the whole set $V(G) \cup V(H)$. As soon as the color partition of $V(G) \cup V(H)$ stabilizes[1], CR terminates and outputs the current coloring $C = C_t$ of $V(G) \cup V(H)$. The color names are renamed for both graphs synchronously.

We say that CR *distinguishes* G and H if $\{\!\{C(x)\}\!\}_{x \in V(G)} \neq \{\!\{C(x)\}\!\}_{x \in V(H)}$. A graph G is called *CR-identifiable* if it is distinguishable by CR from every non-isomorphic H. Note that every CR-discrete graph is CR-identifiable.

3.2 Proof of Theorem 3

3.2.1 Parts 1 and 2

Parts 1 and 2 of Theorem 3 follow immediately from the lemma below. We prove this lemma by a direct combinatorial argument. Alternatively, one can use an algebraic approach in [11, Theorem 2] or the connection to finite variable logics exploited in [4, Lemma 4].

▶ **Lemma 8.** *Let G and H be uncolored n-vertex graphs (the case $G = H$ is not excluded). Let $x \in V(G)$, $x' \in V(H)$, and k be an arbitrary non-negative integer. Then $C_k(x) \neq C_k(x')$ whenever $w_k^G(x) \neq w_k^H(x')$.*

Proof. Using the induction on k, we prove that $w_k^G(x) = w_k^H(x')$ whenever $C_k(x) = C_k(x')$. In the base case of $k = 0$, these equalities are equivalent just because they are both true by definition (recall that $w_0^G(x) = 0$). Assume that $C_k(y) = C_k(y')$ implies $w_k^G(y) = w_k^H(y')$ for all $y \in V(G)$ and $y' \in V(H)$. Let $C_{k+1}(x) = C_{k+1}(x')$. By the definition of the refinement step, we have $\{\!\{C_k(y)\}\!\}_{y \in N(x)} = \{\!\{C_k(y)\}\!\}_{y \in N(x')}$. Using the induction assumption, from here we derive the equality $\{\!\{w_k^G(y)\}\!\}_{y \in N(x)} = \{\!\{w_k^H(y)\}\!\}_{y \in N(x')}$. The equality $w_{k+1}^G(x) = w_{k+1}^H(x')$ now follows by noting that $w_{k+1}^G(x) = \sum_{y \in N(x)} w_k^G(y)$. ◀

3.2.2 Part 3

We now construct a graph G with the three desired properties (a)–(c). Note that this graph can be used in an obvious way to produce infinitely many examples separating the strength of WM and CR.

Let $\mathbb{Z}_n$ denote the cyclic group with elements $0, 1, \ldots, n$ and operation being the addition modulo n. Our construction is based on the well-known Shrikhande graph; see, e.g., [12]. This is the Cayley graph of the group $\mathbb{Z}_4 \times \mathbb{Z}_4$ with connection set $\{\pm(1,0), \pm(0,1), \pm(1,1)\}$. A natural drawing of the Shrikhande graph on the torus can be seen in both parts of Fig. 1.

Recall that a graph G is *strongly regular* with parameters (n, d, λ, μ) if it has n vertices, every vertex in G has d neighbors (i.e., G is *regular* of degree d), every two adjacent vertices of G have λ common neighbors, and every two non-adjacent vertices have μ common neighbors. We will use two properties of the Shrikhande graph:

- It is a strongly regular graph with parameters $(16, 6, 2, 2)$.
- The pairs u, v of non-adjacent vertices in the graph are split into two categories depending on whether the two common neighbors of u and v are adjacent or not.

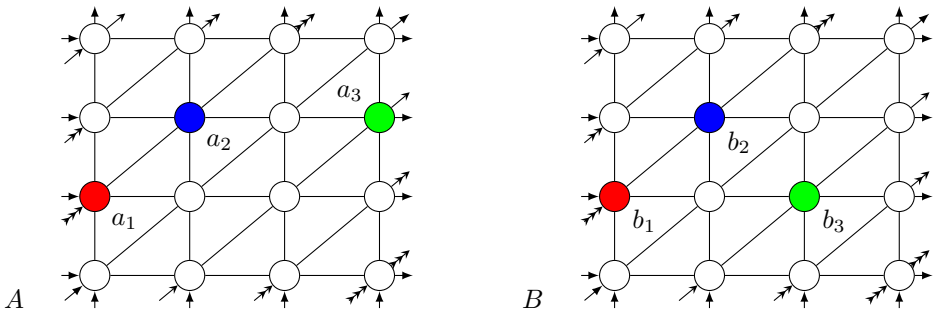

Figure 1 Two colored versions of the Shrikhande graph.

Consider two copies A and B of the Shrikhande graph. In each of A and B, let us individualize three vertices, a_1, a_2, a_3 in A and b_1, b_2, b_3 in B, by assigning unique colors as shown in Fig. 1. The vertices a_i and b_i for each $i = 1, 2, 3$ are equally colored. All non-individualized vertices are considered also colored, all in the same color. Of the three vertices a_1, a_2, a_3, only a_1 and a_2 are adjacent, and the vertices b_1, b_2, b_3 have the same adjacency pattern. An important difference between A and B is that the two common neighbors of b_2 and b_3 are adjacent while the two common neighbors of a_2 and a_3 are not.[2] This implies that the vertex-colored graphs A and B are non-isomorphic.

Before presenting further details, we give a brief outline of the rest of the proof. We will begin with establishing some useful properties of A and B. Though these graphs are non-isomorphic, it is useful to notice that they are still quite similar in the sense that they are indistinguishable by one round of CR (Claim 9). On the other hand, both A and B are CR-discrete (Claim 10) and are, therefore, distinguished after CR makes sufficiently many rounds (Claim 11). The desired graph G will be constructed from A and B by connecting the equally colored vertices, i.e., a_i and b_i, via new edges and vertices. While $a_1, a_2, a_3, b_1, b_2, b_3$ are not colored any more in G, their neighborhoods are modified so that their colors are actually simulated by iterated degrees. This allows us to derive from Claims 10 and 11 that G is CR-discrete (Claim 12). On the other hand, G is not WM-discrete (Claim 14). In order to show that some vertices in G have the same numbers of outgoing walks of each length, we use some basic properties of strongly regular graphs (Claim 13) and the fact that a walk can leave A or B only via one of the vertices $a_1, a_2, a_3, b_1, b_2, b_3$ (and here an important role is played by Claim 9). Finally, we argue that G is not WM-identifiable (Claim 15). Indeed, if we construct another graph G' similarly to G but using two copies of A, then G and G' will have the same walk matrix.

We now proceed with the detailed proof.

▷ Claim 9. After the first round of CR, the vertex-colored graphs A and B are still indistinguishable. That is, there is a bijection $f : V(A) \to V(B)$ such that $C_1(x) = C_1(f(x))$ for all $x \in V(A)$.

[1] Note that the stabilization on each of the sets $V(G)$ and $V(H)$ can occur earlier than on $V(G) \cup V(H)$.
[2] Using the fact that the Shrikhande graph is arc-transitive, it is easy to check that A and B are defined uniquely up to isomorphism of colored graphs.

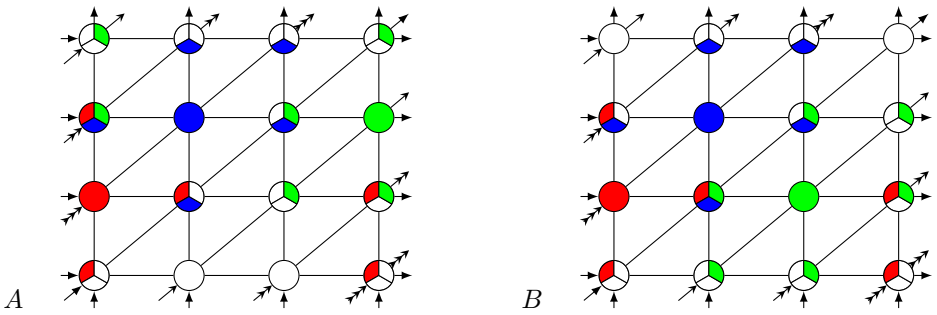

Figure 2 The colorings of A and B after the first color refinement round. For each $i = 1, 2, 3$, the vertices a_i and b_i have the same unique color. The color of each non-individualized vertex is determined by its adjacency to the individualized vertices. For example, the color of a vertex in A means that this vertex is adjacent to a_1 and a_2 but not to a_3.

Proof. Since A and B are regular graphs of the same degree, the equally colored vertices $x \in V(A)$ and $x' \in V(B)$ obtain distinct colors after the first color refinement round only when their neighborhoods contain different sets of individualized vertices (that is, $a_i \in N(x)$ while $b_i \notin N(x')$ or vice versa for some $i = 1, 2, 3$). This is not the case for the individualized vertices because the correspondence $a_i \mapsto b_i$ is a partial isomorphism of A and B. As for the non-individualized vertices, both A and B have exactly one vertex adjacent to all the three individualized vertices, three vertices adjacent to exactly two of them, and two vertices adjacent to none of them. Moreover, in both A and B there are two non-individualized vertices adjacent to a_i (resp. to b_i) for each $i = 1, 2$ and three non-individualized vertices adjacent to a_3 (resp. to b_3). The colorings of A and B after the first refinement round are shown in Fig. 2. ◁

▷ Claim 10. Both vertex-colored graphs A and B are CR-discrete.

Proof. Call a vertex *solitary* if CR colors it differently than the other vertices of the graphs. We prove that every vertex in A is solitary. Virtually the same argument applies also to B. The individualized vertices a_1, a_2, a_3 are solitary from the very beginning. The single vertex a adjacent to all of them is obviously also solitary. Thus, A contains a triangle subgraph whose all vertices, namely a, a_1, a_2, are solitary. Let a' be the common neighbor of a_1 and a_2 different from a (recall that the Shrikhande graph is strongly regular with the third parameter $\lambda = 2$). The fact that a_1 and a_2 are solitary implies that the equality $C(a') = C(x)$ for $x \neq a'$ can be true only if $x = a$, which is actually impossible because a is solitary. Therefore, a' is solitary too. This argument applies to any triangle whose all vertices are solitary and to the other common neighbor of any two vertices of this triangle. Consider the graph whose vertices are the triangles of the Shrikhande graph, adjacent exactly when they share an edge. This graph (known as the *Dyck graph*) is obviously connected, which readily implies that all vertices of A are solitary. ◁

▷ Claim 11. The vertex-colored graphs A and B are distinguishable by CR.

Proof. Recall that A and B are non-isomorphic because the two common neighbors of b_2 and b_3 are adjacent while the two common neighbors of a_2 and a_3 are not. By Claim 10, both A and B are CR-discrete. Assume that A and B are indistinguishable by CR. Let f be

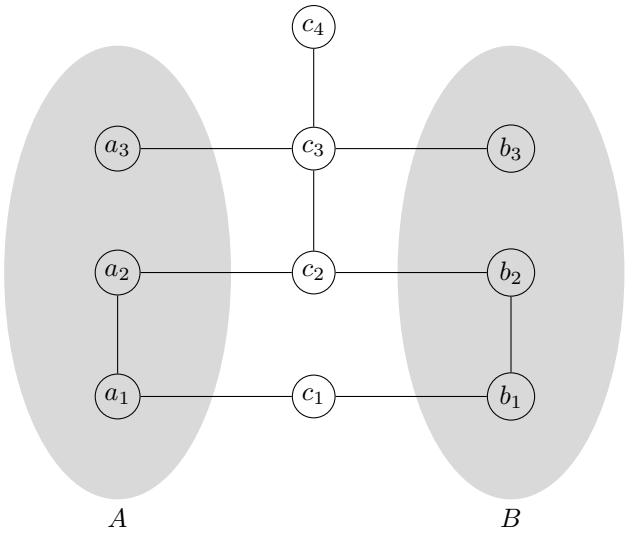

Figure 3 Construction of G.

the bijection from $V(A)$ to $V(B)$ respecting the final CR-coloring, that is, $C(f(x)) = C(x)$ for all $x \in V(A)$. Since f is not an isomorphism, there are vertices u and v in A such that u and v are adjacent but $f(u)$ and $f(v)$ are not or vice versa. This shows, however, that the coloring C is still unstable because, by the refinement rule, u and $f(u)$ have to receive distinct colors in the next round. This contradiction proves the claim. ◁

We now construct a graph G as the vertex-disjoint union of A and B where each pair a_i, b_i is connected via new edges and new intermediate vertices as shown in Fig. 3. Thus, $V(G) = V(A) \cup V(B) \cup \{c_1, c_2, c_3, c_4\}$ where c_1, c_2, c_3, c_4 are new connector vertices of degree $2, 3, 4, 1$ respectively. The graph G is uncolored, that is, the colors of the six individualized vertices $a_1, a_2, a_3, b_1, b_2, b_3$ are erased. The next claim proves Part 3(a) of the theorem.

▷ Claim 12. G is CR-discrete.

Proof. The connector vertices c_1, c_2, c_3, c_4 have unique degrees $2, 3, 4, 1$ and become solitary after the first refinement round. The vertices a_1, a_2, a_3 have degree 7, while the other vertices in A have degree 6. Each of the three vertices a_1, a_2, a_3 is distinguished from the other two by the adjacency to its own connector. It follows that after the second refinement round, the colors $C_2(a_1), C_2(a_2), C_2(a_3)$ become unique within A (even when still $C_2(a_i) = C_2(b_i)$). Claim 10, therefore, implies that eventually $C(x) \neq C(x')$ for all $x \neq x'$ in A. The same argument applies to B. Using the same argument as in the proof of Claim 11, we also have $C(x) \neq C(x')$ for all $x \in V(A)$ and $x' \in V(B)$. ◁

Let $w_k^R(x, y)$ denote the number of walks of length k from a vertex x to a vertex y in a graph R. We will need the following simple and well-known facts.[3]

[3] Let P_{k+1} be a path of length k with end vertices s and t. Note that $w_k^R(x, y)$ is equal to the number of all homomorphisms from P_{k+1} to R taking s to x and t to y. Part 2 of Claim 13 is a particular case of a much more general result about the invariance of homomorphism counts under the Weisfeiler-Leman equivalence for graphs with designated vertices [4, Lemma 4].

▷ Claim 13.

1. If R is a regular graph of degree d, then $w_k^R(x) = d^k$ for every $x \in V(R)$.
2. Suppose now that R is a strongly regular graph with parameters (n, d, λ, μ) and fix an arbitrary $k \geq 0$. Then the walk count $w_k^R(x, x)$ is the same for every $x \in V(R)$. If $x \neq y$, then the value of $w_k^R(x, y)$ depends only on the adjacency of x and y (and on the parameters d, λ, μ).

Proof. Part 1 is obvious. Part 2 follows from an easy inductive argument. Indeed, it is trivially true for $k = 0$. Assume that $w_k^R(x, y) = a_k$ for all adjacent x and y and that $w_k^R(x, y) = n_k$ for all non-adjacent unequal x and y. Then $w_{k+1}^R(x, x) = \sum_{z \in N(x)} w_k^R(z, x) = d\, a_k$. If x and y are adjacent, then

$$w_{k+1}^R(x, y) = \sum_{z \in N(x) \cap N(y)} w_k^R(z, y) + \sum_{z \in N(x) \setminus N(y)} w_k^R(z, y) = \lambda a_k + (d - \lambda) n_k.$$

If x and y are non-adjacent and unequal, then

$$w_{k+1}^R(x, y) = \sum_{z \in N(x) \cap N(y)} w_k^R(z, y) + \sum_{z \in N(x) \setminus N(y)} w_k^R(z, y) = \mu a_k + (d - \mu) n_k,$$

enabling the induction step. ◁

We are now prepared to prove Part 3(b) of the theorem.

▷ Claim 14. G is not WM-discrete.

Proof. Define an equivalence relation $\equiv$ on $V(G)$ as follows. Each connector vertex is equivalent only to itself. Let C_1 be the coloring of $V(A) \cup V(B)$ obtained after the first round of the execution of CR on the vertex-colored graphs A and B; see Claim 9. We set $x \equiv x'$ for $x, x' \in V(A) \cup V(B)$ if $C_1(x) = C_1(x')$. Recall that the largest equivalence class of $\equiv$ consists of six vertices (three uncolored vertices in A adjacent to a_3 but neither to a_1 nor to a_2 and three uncolored vertices in B adjacent to b_3 but neither to b_1 nor to b_2). We claim that $w_k^G(x) = w_k^G(x')$ for every k whenever $x \equiv x'$. Indeed, if $x \in V(A)$, then

$$w_k^G(x) = w_k^A(x) + \sum_{i=1}^{3} \sum_{j=0}^{k-1} w_j^A(x, a_i) w_{k-j-1}^G(c_i). \tag{11}$$

Here, we separately consider the walks of length k inside A and the walks of length k leaving A. A walk can leave A only after visiting one of the vertices a_1, a_2, a_3. If such a walk leaves A first after the j-th step via a_i, it arrives at the connector c_i and, starting from it, makes the remaining $k - j - 1$ steps. The similar equality holds for $x \in V(B)$.

It remains to notice that the right hand side of (11) and its analog for B yield the same value for all x in the same $\equiv$-class. Indeed, let $x \equiv x'$ and suppose that $x \in A$ and $x' \in B$ (the cases $x, x' \in A$ and $x, x' \in B$ are completely similar). Then $w_k^A(x) = w_k^B(x') = 6^k$ by Part 1 of Claim 13. Finally, for each j the equalities $w_j^A(x, a_i) = w_j^B(x, b_i)$ for $i = 1, 2, 3$ follow from Part 2 of Claim 13 by the definition of the relation $\equiv$ and the description of C_1 in the proof of Claim 9. ◁

It remains to prove Part 3(c) of the theorem.

▷ Claim 15. G is not WM-identifiable.

Proof. Construct G' in the same way as G but using a copy A' of A instead of B. The graphs G and G' are non-isomorphic, basically because A and B are non-isomorphic as colored graphs. In particular, G' has an automorphism fixing the connector vertices and transposing A and A', whereas G has no non-trivial automorphism by Claim 12. Fix a colored-graph isomorphism f' from A' to A. Define a bijection F from $V(G') = V(A) \cup V(A') \cup \{c_1, c_2, c_3, c_4\}$ onto $V(G) = V(A) \cup V(B) \cup \{c_1, c_2, c_3, c_4\}$ so that $F(c_i) = c_i$ for $i = 1, 2, 3, 4$, the restriction f of F to $V(A)$ is as in Claim 9, and the restriction of F to $V(A')$ is the isomorphism f'. The proof of Claim 14 applies to the graph G' virtually without changes. In particular, the analog of Equality (11) for G' allows us to show by a simple induction that $w_k^{G'}(x) = w_k^G(f(x))$ for $x \in V(A)$ and $w_k^{G'}(x') = w_k^G(f'(x'))$ for $x' \in V(A')$, as well as that $w_k^{G'}(c_i) = w_k^G(c_i)$ for $i = 1, 2, 3, 4$. Thus, for every $x \in V(G')$ we have $w_k^{G'}(x) = w_k^G(F(x))$ for all k, implying that G and G' are WM-indistinguishable. ◁

The proof of Theorem 3 is complete.

References

1. Noga Alon and Joel H. Spencer. *The probabilistic method*. John Wiley & Sons, 2016.
2. László Babai, Paul Erdős, and Stanley M. Selkow. Random graph isomorphism. *SIAM Journal on Computing*, 9(3):628–635, 1980.
3. László Babai and Ludek Kucera. Canonical labelling of graphs in linear average time. In *20th Annual Symposium on Foundations of Computer Science (FOCS'79)*, pages 39–46. IEEE Computer Society, 1979. doi:10.1109/SFCS.1979.8.
4. Z. Dvořák. On recognizing graphs by numbers of homomorphisms. *Journal of Graph Theory*, 64(4):330–342, 2010.
5. Chris Godsil. Controllable subsets in graphs. *Ann. Comb.*, 16(4):733–744, 2012. doi:10.1007/s00026-012-0156-3.
6. Elias M. Hagos. Some results on graph spectra. *Linear Algebra Appl.*, 356(1-3):103–111, 2002. doi:10.1016/S0024-3795(02)00324-5.
7. David Harvey and Joris van der Hoeven. Integer multiplication in time $O(n \log n)$. *Annals of Mathematics*, 193(2):563–617, 2021. doi:10.4007/annals.2021.193.2.4.
8. Fenjin Liu and Johannes Siemons. Unlocking the walk matrix of a graph. *J. Algebr. Comb.*, 55(3):663–690, 2022. doi:10.1007/s10801-021-01065-3.
9. H. L. Morgan. The generation of a unique machine description for chemical structures — a technique developed at chemical abstracts service. *J. Chem. Doc.*, 5(2):107–113, 1965. doi:10.1021/c160017a018.
10. Sean O'Rourke and Behrouz Touri. On a conjecture of Godsil concerning controllable random graphs. *SIAM J. Control. Optim.*, 54(6):3347–3378, 2016. doi:10.1137/15M1049622.
11. David L. Powers and Mohammad M. Sulaiman. The walk partition and colorations of a graph. *Linear Algebra Appl.*, 48:145–159, 1982. doi:10.1016/0024-3795(82)90104-5.
12. Sharad S. Sane. The Shrikhande graph. *Resonance*, 20:903–918, 2015. doi:10.1007/s12045-015-0255-7.