Check for updates

J. Appl. Probab. **61,** 1139–1152 (2024)
*doi:*10.1017/jpr.2023.112

AVERAGE JACCARD INDEX OF RANDOM GRAPHS

QUNQIANG FENG ,*
SHUAI GUO,* AND
ZHISHUI HU ,* ** *University of Science and Technology of China*

Abstract

The asymptotic behavior of the Jaccard index in $G(n, p)$, the classical Erdös–Rényi random graph model, is studied as n goes to infinity. We first derive the asymptotic distribution of the Jaccard index of any pair of distinct vertices, as well as the first two moments of this index. Then the average of the Jaccard indices over all vertex pairs in $G(n, p)$ is shown to be asymptotically normal under an additional mild condition that $np \to \infty$ and $n^2(1-p) \to \infty$.

Keywords: Erdös–Rényi Random graph; Jaccard similarity; asymptotic distribution; inverse moment

2020 Mathematics Subject Classification: Primary 05C80; 60C05
Secondary 60F05

1. Introduction

The *Jaccard index*, also known as the *Jaccard similarity coefficient*, was originally introduced by Paul Jaccard to measure the similarity between two sets [13]. For any two finite sets $\mathcal{A}$ and $\mathcal{B}$, the Jaccard index $J(\mathcal{A}, \mathcal{B})$ is the ratio of the size of their intersection to the size of their union. That is,

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A}| + |\mathcal{B}| - |\mathcal{A} \cap \mathcal{B}|},$$

where the symbol $|\cdot|$ denotes the cardinality of a set. It is clear that this index ranges from 0 to 1. The associated *Jaccard distance* for quantifying the dissimilarity between two sets is defined as one minus the Jaccard index (see, e.g., [11, 17]). In statistics and data science, the Jaccard index is employed as a statistic to measure the similarity between sample sets, especially for binary and categorical data (see, e.g., [6, 15]). For extensive generalizations of the Jaccard index in many other mathematical structures, such as scalars, vectors, matrices, and multisets, we refer the reader to [7]. Due to its simplicity and popularity, many applications of the Jaccard index and its variants have been developed in various fields, such as cell formation [29], pattern recognition [12], data mining [24], natural language processing [27], recommendation systems [3], medical image segmentation [8], and machine learning [1].

Following the original definition on sets, the Jaccard index of two vertices in a graph can naturally be extended to equal the number of common neighbors divided by the number of

Received 5 June 2023; accepted 17 November 2023.
* Postal address: Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China.
** Email address: huzs@ustc.edu.cn

© The Author(s), 2024. Published by Cambridge University Press on behalf of Applied Probability Trust.

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

vertices that are neighbors of at least one of them (see, e.g., [9]). As a graph benchmark suitable for real-world applications, the Jaccard index has also been proposed to determine the similarity in graphs or networks [16], because of its clear interpretability and computational scalability. This index, as well as its variants, is employed to find core nodes for community detection in complex networks [4, 20], to estimate the coupling strength between temporal graphs [19], and to do link prediction [18, 21, 30], among others.

Erdös–Rényi random graphs are widely used as a benchmark model in statistical network analysis (see, e.g., [2, 26]). In the simulation study of [22], it is shown that the empirical cumulative distribution functions of Jaccard indices over all vertex pairs in two network models, the Erdös–Rényi random graph and the stochastic block model, are quite different. Despite the widespread applications of the Jaccard index in network analysis, to the best of our knowledge there is a lack of comprehensive study of theoretical results on this simple index defined on statistical graph models. As the first step toward filling this gap, our main concern in this paper is derive the asymptotic behavior of the basic Jaccard index in Erdös–Rényi random graphs. For numerous probabilistic results on this classical random graph model we refer the reader to [5, 14, 25].

Throughout this paper we use the following notation. For any integer $n \geq 2$, we denote by $[n]$ the vertex set $\{1, 2, \ldots, n\}$. For an event $\mathcal{E}$, let $|\mathcal{E}|$ be the cardinality, $\overline{\mathcal{E}}$ the complement, and $\mathbf{1}(\mathcal{E})$ the indicator of $\mathcal{E}$. For real numbers a, b, we write $a \vee b$ to denote the maximum of a and b. For probabilistic convergence, we use $\xrightarrow{\mathrm{D}}$ and $\xrightarrow{\mathrm{P}}$ to denote convergence in distribution and in probability, respectively.

The rest of this paper is organized as follows. The Jaccard index of any pair of distinct vertices in the Erdös–Rényi random graph $\mathrm{G}(n, p)$ is considered in Section 2. We first compute the mean and variance of this index, and then obtain the phase changes of its asymptotic distribution for $p \in [0, 1]$ in all regimes as $n \to \infty$. In Section 3, we prove the asymptotic normality of the average of the Jaccard indices over all vertex pairs in $G(n, p)$ as $np \to \infty$ and $n^2(1-p) \to \infty$.

2. Jaccard index of a vertex pair

Let us denote by $G(n, p)$ an Erdös–Rényi random graph on the vertex set $[n]$, where each edge is present independently with probability p. In this paper we consider $p = p(n)$ as a function of the graph size n. For any two vertices $i, j \in [n]$, let $\mathbf{1}_{ij}$ be the indicator that takes the value 1 if an edge between i and j is present in $G(n, p)$, and takes the value 0 otherwise. It follows that $\mathbf{1}_{ii} = 0$, $\mathbf{1}_{ij} = \mathbf{1}_{ji}$, and $\{\mathbf{1}_{ij} : 1 \leq i < j \leq n\}$ is a sequence of independent Bernoulli variables with success rate p. The $n \times n$ matrix $\boldsymbol{A} = (\mathbf{1}_{ij})$ is usually called the adjacent matrix of $G(n, p)$, and is a symmetric matrix with all diagonal entries equal to zero.

For any vertex $i \in [n]$, let the set $\mathcal{N}_i$ be its neighborhood, i.e., $\mathcal{N}_i = \{k : \mathbf{1}_{ik} = 1, k \in [n]\}$. For any pair of vertices $i, j \in [n]$, we also define their union neighborhood as

$$\mathcal{N}_{ij} = \{k : \mathbf{1}_{ik} \vee \mathbf{1}_{jk} = 1,\ k \in [n] \text{ and } k \neq i, j\}, \quad i \neq j.$$

Notice that here the neighborhood set $\mathcal{N}_{ij}$ does not contain vertices i and j themselves, even if $\mathbf{1}_{ij} = 1$. Then the Jaccard index of vertices i and j in $G(n, p)$ is formally defined as

$$J_{ij}^{(n)} = \frac{|\mathcal{N}_i \cap \mathcal{N}_j|}{|\mathcal{N}_{ij}|} =: \frac{S_{ij}^{(n)}}{T_{ij}^{(n)}}, \quad i \neq j. \tag{1}$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

We can see that the index $J_{ij}^{(n)}$ given in (1) is not well defined when $\mathcal{N}_{ij}$ is an empty set or $T_{ij}^{(n)} = 0$. For convenience, following the idea in [6], we define $J_{ij}^{(n)} = p/(2-p)$ in this special case. Indeed, it is shown later that the conditional expectation of $J_{ij}^{(n)}$ is exactly $p/(2-p)$ given that $T_{ij}^{(n)} > 0$. In terms of the adjacent matrix $\boldsymbol{A}$, the numerator and denominator in (1) can be rewritten as

$$S_{ij}^{(n)} = \sum_{k \neq i,j} \mathbf{1}_{ik}\mathbf{1}_{jk}, \qquad T_{ij}^{(n)} = \sum_{k \neq i,j} \mathbf{1}_{ik} \vee \mathbf{1}_{jk}. \tag{2}$$

Due to the independence of the elements in $\boldsymbol{A}$, it is clear that the random variables $S_{ij}^{(n)}$ and $T_{ij}^{(n)}$ follow the binomial distributions $\mathrm{Bin}(n-2, p^2)$ and $\mathrm{Bin}(n-2, p(2-p))$, respectively. Hence, the Jaccard index of a vertex pair in $G(n, p)$ is a quotient of two dependent binomial random variables.

2.1. Mean and variance

We first calculate the mean and variance of the Jaccard index of any pair of vertices in G. By (1) and (2), we can see that $\left\{J_{ij}^{(n)},\ 1 \le i < j \le n\right\}$ is a sequence of random variables that are pairwise dependent but identically distributed. Without loss of generality, we only consider $J_{12}^{(n)}$.

For any vertex $3 \le k \le n$, the conditional probability

$$\mathbb{P}(\mathbf{1}_{1k}\mathbf{1}_{2k} = 1 \mid \mathbf{1}_{1k} \vee \mathbf{1}_{2k} = 1) = \frac{\mathbb{P}(\mathbf{1}_{1k}\mathbf{1}_{2k} = 1)}{\mathbb{P}(\mathbf{1}_{1k} \vee \mathbf{1}_{2k} = 1)} = \frac{p}{2-p},$$

which is independent of k. Then, for any positive integer $1 \le m \le n-2$, given the event $T_{12}^{(n)} = m$, the conditional distribution of $S_{12}^{(n)}$ is $\mathrm{Bin}(m, p/(2-p))$, due to independence of the indicators $\{\mathbf{1}_{1k}, \mathbf{1}_{2k},\ 3 \le k \le n\}$. Consequently, we have

$$\mathbb{E}\left[S_{12}^{(n)} \mid T_{12}^{(n)} = m\right] = \frac{mp}{2-p}, \tag{3}$$

$$\mathrm{Var}\left[S_{12}^{(n)} \mid T_{12}^{(n)} = m\right] = \frac{2mp(1-p)}{(2-p)^2}. \tag{4}$$

Noting that $J_{12}^{(n)} = p/(2-p)$ in the special case $T_{12}^{(n)} = 0$, by (1) and (3) we thus have $\mathbb{E}\left[J_{12}^{(n)} \mid T_{12}^{(n)}\right] = p/(2-p)$, which implies that $\mathbb{E}\left[J_{12}^{(n)}\right] = p/(2-p)$ and $\mathrm{Var}\left[\mathbb{E}\left(J_{12}^{(n)} \mid T_{12}^{(n)}\right)\right] = 0$. Using the law of total variance, it follows from this and (4) that

$$\begin{aligned}\mathrm{Var}\left[J_{12}^{(n)}\right] = \mathbb{E}\left[\mathrm{Var}\left(J_{12}^{(n)} \mid T_{12}^{(n)}\right)\right] &= \sum_{m=1}^{n-2} \mathbb{P}\left(T_{12}^{(n)} = m\right)\mathrm{Var}\left[\frac{S_{12}^{(n)}}{m} \mid T_{12}^{(n)} = m\right] \\ &= \frac{2p(1-p)}{(2-p)^2} \sum_{m=1}^{n-2} \frac{1}{m}\mathbb{P}\left(T_{12}^{(n)} = m\right),\end{aligned} \tag{5}$$

which involves the first inverse moment of the binomial distribution. Recalling that $T_{12}^{(n)}$ has the distribution $\mathrm{Bin}(n-2, p(2-p))$, it follows by [28, Corollary 1] that

$$\sum_{m=1}^{n-2} \frac{1}{m}\mathbb{P}\left(T_{12}^{(n)} = m\right) = \frac{1}{np(2-p)}\left(1 + O\left(\frac{1}{np}\right)\right) \tag{6}$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

as $np \to \infty$. By (5) and (6), we thus have

$$\mathrm{Var}\left[J_{12}^{(n)}\right] = \frac{2(1-p)}{n(2-p)^3}\left(1 + O\left(\frac{1}{np}\right)\right).$$

Collecting the above findings, by Chebyshev's inequality we thus prove the following result.

Proposition 1. *Let $J_{ij}^{(n)}$ be the Jaccard index of any distinct vertices $i, j \in [n]$ in G(n,p). Then $\mathbb{E}\left[J_{ij}^{(n)}\right] = p/(2-p)$ for all $n \geq 2$. In particular, as $np \to \infty$, it further follows that*

$$\mathrm{Var}\left[J_{ij}^{(n)}\right] = \frac{2(1-p)}{n(2-p)^3}\left(1 + O\left(\frac{1}{np}\right)\right),$$

and $J_{ij}^{(n)} - p/(2-p)$ converges to 0 in probability.

2.2. Asymptotic distribution

We now establish the asymptotic distribution of the Jaccard index of any vertex pair in $G(n, p)$.

Theorem 1. *Let $J_{ij}^{(n)}$ be the Jaccard index of any distinct vertices $i, j \in [n]$ in G(n,p).*

(i) *If $np^2(1-p) \to \infty$, then*

$$\sqrt{\frac{n(2-p)^3}{2(1-p)}}\left(J_{ij}^{(n)} - \frac{p}{2-p}\right) \xrightarrow{\mathrm{D}} Z,$$

where Z denotes a standard normal random variable.

(ii) *If $np^2 \to \lambda$ for some constant $\lambda > 0$, then $2npJ_{ij}^{(n)} \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda)$, where $\mathrm{Poi}(\lambda)$ denotes the Poisson distribution with parameter λ.*

(iii) *If $np^2 \to 0$, then $npJ_{ij}^{(n)} \xrightarrow{\mathrm{P}} 0$.*

(iv) *If $n(1-p) \to c$ for some constant $c > 0$, then $n\left(1 - J_{ij}^{(n)}\right) \xrightarrow{\mathrm{D}} \mathrm{Poi}(2c)$.*

(v) *If $n(1-p) \to 0$, then $n\left(1 - J_{ij}^{(n)}\right) \xrightarrow{\mathrm{P}} 0$.*

Proof. As presented in the previous subsection, it is sufficient to consider a single index $J_{12}^{(n)}$. To prove (i), we first rewrite

$$\begin{aligned}\sqrt{\frac{(n-2)(2-p)^3}{2(1-p)}}\left(J_{12}^{(n)} - \frac{p}{2-p}\right) &= \sqrt{\frac{(n-2)(2-p)}{2(1-p)}} \cdot \frac{(2-p)S_{12}^{(n)} - pT_{12}^{(n)}}{T_{12}^{(n)}} \\ &= \frac{(2-p)S_{12}^{(n)} - pT_{12}^{(n)}}{\sqrt{2(n-2)p^2(1-p)(2-p)}} \cdot \frac{(n-2)p(2-p)}{T_{12}^{(n)}}. \qquad (7)\end{aligned}$$

For any distinct vertices $i, j, k \in [n]$, we write

$$V_{ij,k} = (2-p)\mathbf{1}_{ik}\mathbf{1}_{jk} - p(\mathbf{1}_{ik} \vee \mathbf{1}_{jk}). \qquad (8)$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

Then, for any fixed two vertices $i, j \in [n]$, the random variables $\{V_{ij,k},\ k \in [n], k \neq i, j\}$ are independent and identically distributed with common mean 0 and common variance $2p^2(1-p)(2-p)$. Since it follows by (2) that

$$(2-p)S_{12}^{(n)} - pT_{12}^{(n)} = \sum_{k=3}^{n} V_{12,k}, \tag{9}$$

a direct application of the Lindeberg–Feller central limit theorem yields

$$\frac{(2-p)S_{12}^{(n)} - pT_{12}^{(n)}}{\sqrt{2(n-2)p^2(1-p)(2-p)}} \xrightarrow{\mathrm{D}} Z \tag{10}$$

whenever $np^2(1-p) \to \infty$. By Chebyshev's inequality, the fact that $T_{12}^{(n)} \sim \mathrm{Bin}((n-2), p(2-p))$ gives us that, as $np \to \infty$,

$$\frac{T_{12}^{(n)}}{(n-2)p(2-p)} \xrightarrow{\mathrm{P}} 1, \tag{11}$$

which, together with (7) and (10), proves (i) by Slutsky's lemma.

If $np^2 \to \lambda$ for some constant $\lambda > 0$, we must have that $p \to 0$ and $np \to \infty$. Since $S_{12}^{(n)} \sim \mathrm{Bin}(n-2, p^2)$, the Poisson limit theorem yields that $S_{12}^{(n)} \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda)$. By (11), again using Slutsky's lemma, we obtain

$$2npJ_{12}^{(n)} = \frac{2np}{T_{12}^{(n)}} \cdot S_{12}^{(n)} \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda),$$

which proves (ii).

If $np^2 \to 0$, to prove (iii) we only need to prove that the probability $\mathbb{P}\Big(npJ_{ij}^{(n)} > np^2/(2-p)\Big)$ tends to 0. Note that if the event $\left\{S_{12}^{(n)} = 0\right\}$ occurs, the Jaccard index $J_{ij}^{(n)}$ must be 0 or $p/(2-p)$. Therefore,

$$\mathbb{P}\left(npJ_{ij}^{(n)} > \frac{np^2}{2-p}\right) = \mathbb{P}\left(J_{ij}^{(n)} > \frac{p}{2-p}\right) \le \mathbb{P}\left(S_{12}^{(n)} \neq 0\right) = 1 - \mathbb{P}\left(S_{12}^{(n)} = 0\right).$$

Again by the fact that $S_{12}^{(n)} \sim \mathrm{Bin}(n-2, p^2)$, we have $1 - \mathbb{P}\big(S_{12}^{(n)} = 0\big) \to 0$ if $np^2 \to 0$. This implies (iii).

Analogously to (7), we have

$$\frac{p}{2-p} - J_{12}^{(n)} = \frac{-(2-p)S_{12}^{(n)} + pT_{12}^{(n)}}{(n-2)p(2-p)^2} \cdot \frac{(n-2)p(2-p)}{T_{12}^{(n)}}. \tag{12}$$

Note that (11) still holds, and $n[1-p/(2-p)]$ has the limit $2c$ if $n(1-p) \to c$ for some constant $c > 0$. To prove (iv), by (12) it is now sufficient to show that

$$-(2-p)S_{12}^{(n)} + pT_{12}^{(n)} \xrightarrow{\mathrm{D}} \mathrm{Poi}(2c) - 2c. \tag{13}$$

For any distinct vertices $i, j, k \in [n]$, we can directly obtain from (8) that the characteristic function of $-V_{ij,k}$ is

$$f_n(t) = \mathbb{E}\big[\mathrm{e}^{-\mathrm{i}tV_{ij,k}}\big] = p^2\mathrm{e}^{-2\mathrm{i}t(1-p)} + 2p(1-p)\mathrm{e}^{\mathrm{i}tp} + (1-p)^2,$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

where $\mathrm{i} = \sqrt{-1}$ denotes the imaginary unit. Then, by (9) and independence, we have that the characteristic function of $-(2-p)S_{12}^{(n)} + pT_{12}^{(n)}$ is equal to

$$f_n^{n-2}(t) = \left[p^2 \mathrm{e}^{-2\mathrm{i}t(1-p)} + 2p(1-p)\mathrm{e}^{\mathrm{i}tp} + (1-p)^2\right]^{n-2}, \quad t \in \mathbb{R}.$$

Note that

$$\begin{aligned}\lim_{n\to\infty} n\left[p^2 \mathrm{e}^{-2\mathrm{i}t(1-p)} + 2p(1-p)\mathrm{e}^{\mathrm{i}tp} + (1-p)^2 - 1\right] &= 2c\mathrm{e}^{\mathrm{i}t} + \lim_{n\to\infty} n\left[p^2 \mathrm{e}^{-2\mathrm{i}t(1-p)} - 1\right] \\ &= 2c\mathrm{e}^{\mathrm{i}t} + \lim_{n\to\infty} n\left[p^2\left(\mathrm{e}^{-2\mathrm{i}t(1-p)} - 1\right)\right. \\ &\quad \left. + \left(p^2 - 1\right)\right] \\ &= 2c\left(\mathrm{e}^{\mathrm{i}t} - \mathrm{i}t - 1\right)\end{aligned}$$

if $n(1-p) \to c$. Therefore, the limit of the characteristic function of $-(2-p)S_{12}^{(n)} + pT_{12}^{(n)}$ satisfies

$$\lim_{n\to\infty} f_n^{n-2}(t) = \exp\left\{2c\left(\mathrm{e}^{\mathrm{i}t} - \mathrm{i}t - 1\right)\right\},$$

which implies (13) and completes the proof of (iv).

We only sketch the proof of (v), since it is very similar to (iv). By (11) and (12), it is sufficient to show that $-(2-p)S_{12}^{(n)} + pT_{12}^{(n)}$ converges in probability to 0 under the condition $n(1-p) \to 0$. In fact, following the proof of (13), in this case we can deduce that its characteristic function $f_n^{n-2}(t) \to 1$. □

3. Average Jaccard index

In this section we derive asymptotic properties of the average Jaccard index of $G(n, p)$, which is given by

$$J_n = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} J_{ij}^{(n)}. \tag{14}$$

That is, the average Jaccard index J_n is the average of the Jaccard indices over all vertex pairs in the Erdös–Rényi random graph $\mathrm{G}(n, p)$. An immediate consequence of Proposition 1 is that the expectation of J_n is equal to $p/(2-p)$.

We now state the main results of this paper.

Theorem 2. *Let J_n be the average Jaccard index of G(n,p). If $np \to \infty$ and $n^2(1-p) \to \infty$, then*

$$\frac{n(2-p)^2}{\sqrt{8p(1-p)}}\left(J_n - \frac{p}{2-p}\right) \xrightarrow{\mathrm{D}} Z,$$

where Z denotes a standard normal random variable.

It is remarkable that the quantity $n^2(1-p)/2$ is the asymptotic expected number of unoccupied edges in $G(n, p)$. Theorem 2 suggests that if $G(n, p)$ is neither too sparse (see, e.g., [25, Section 1.8]) nor close to a complete graph, then its average Jaccard index J_n has asymptotic

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

normality. In order to prove Theorem 2, we first introduce two auxiliary lemmas, both of which involve the inverse moments of the binomial distribution.

Lemma 1. *If the random variable X_n has a binomial distribution with parameters n and p, then, for any fixed constants $a \geq 0$ and $b > 0$, as $np \to \infty$,*

$$\mathbb{E}\left[\left(\frac{(n+a)p}{b+X_n} - 1\right)^2\right] = O\left(\frac{1}{np}\right).$$

Proof. For any $\varepsilon \in (0, 1)$, we define $\mathcal{A}_n := \{(1-\varepsilon)(n+a)p \leq b + X_n \leq (1+\varepsilon)(n+a)p\}$. Applying Chernoff's bound for the binomial distribution (see, e.g., [14, Corollary 2.3]) gives, for sufficiently large n,

$$\mathbb{P}(\overline{\mathcal{A}}_n) \leq \mathbb{P}\left(\left|\frac{X_n}{np} - 1\right| > \frac{\varepsilon}{2}\right) \leq 2\exp\left\{-\frac{\varepsilon^2}{12}np\right\} = O\left(\frac{1}{n^3p^3}\right).$$

Then, for sufficiently large n, by noting that

$$\left(\frac{(n+a)p}{b+X_n} - 1\right)^2 \leq \left(\frac{(n+a)p}{b}\right)^2 + 1 \leq 2n^2p^2,$$

we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{(n+a)p}{b+X_n} - 1\right)^2\right] &= \mathbb{E}\left[\left(\frac{(n+a)p}{b+X_n} - 1\right)^2 \mathbf{1}(\mathcal{A}_n)\right] + \mathbb{E}\left[\left(\frac{(n+a)p}{b+X_n} - 1\right)^2 \mathbf{1}(\overline{\mathcal{A}}_n)\right] \\
&\leq \mathbb{E}\left[\frac{(n+a)^2p^2}{(b+X_n)^2}\left(\frac{b+X_n}{(n+a)p} - 1\right)^2 \mathbf{1}(\mathcal{A}_n)\right] + 2n^2p^2\mathbb{P}(\overline{\mathcal{A}}_n) \\
&\leq (1-\varepsilon)^{-2}\mathbb{E}\left[\left(\frac{b+X_n}{(n+a)p} - 1\right)^2\right] + 2n^2p^2\mathbb{P}(\overline{\mathcal{A}}_n) \\
&= O\left(\frac{1}{np}\right),
\end{aligned}$$

where we used the inequality

$$\mathbb{E}\left[\left(\frac{b+X_n}{(n+a)p} - 1\right)^2\right] \leq \frac{2}{(n+a)^2p^2}\left(\mathbb{E}\left[(X_n - np)^2\right] + (b - ap)^2\right) = O\left(\frac{1}{np}\right).$$

□

Lemma 2. *If the random variable X_n has a binomial distribution with parameters n and p, then, for any fixed positive constants b and α, as $np \to \infty$,*

$$\mathbb{E}\left[\frac{1}{(b+X_n)^\alpha}\right] = \frac{1}{(np)^\alpha}(1 + o(1)).$$

Proof. This is an immediate consequence of [23, Theorem 2]. □

We now give a formal proof of Theorem 2.

Proof of Theorem 2. By (12), the index $J_{ij}^{(n)}$ can be expressed as

$$J_{ij}^{(n)} = \frac{p}{2-p} + \frac{(2-p)S_{ij}^{(n)} - pT_{ij}^{(n)}}{(n-2)p(2-p)^2} + R_{ij}^{(n)}, \quad 1 \leq i \neq j \leq n, \tag{15}$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

where the remainder term is

$$R_{ij}^{(n)} = \frac{(2-p)S_{ij}^{(n)} - pT_{ij}^{(n)}}{(n-2)p(2-p)^2}\left(\frac{(n-2)p(2-p)}{T_{ij}^{(n)}} - 1\right). \tag{16}$$

Note the special case in (15) that the remainder term $R_{ij}^{(n)}$ vanishes if $T_{ij}^{(n)} = 0$. Taking expectation on both sides of (15) gives, for any distinct vertices $i, j \in [n]$,

$$\mathbb{E}\left[R_{ij}^{(n)}\right] = 0. \tag{17}$$

Denote by R_n the sum of all the remainder terms, i.e.,

$$R_n = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} R_{ij}^{(n)}. \tag{18}$$

Then it follows by (17) that $\mathbb{E}[R_n] = 0$. By (2) and the simple fact that $\mathbf{1}_{ik} \vee \mathbf{1}_{jk} = \mathbf{1}_{ik} + \mathbf{1}_{jk} - \mathbf{1}_{ik}\mathbf{1}_{jk}$, we have, for any $1 \le i \ne j \le n$,

$$(2-p)S_{ij}^{(n)} - pT_{ij}^{(n)} = \sum_{k\ne i,j}\left[(2-p)\mathbf{1}_{ik}\mathbf{1}_{jk} - p\mathbf{1}_{ik} \vee \mathbf{1}_{jk}\right] = \sum_{k\ne i,j}\left[2\mathbf{1}_{ik}\mathbf{1}_{jk} - p\left(\mathbf{1}_{ik} + \mathbf{1}_{jk}\right)\right],$$

which, together with (14) and (15), implies that

$$\begin{aligned} J_n &= \frac{p}{2-p} + \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{(2-p)S_{ij}^{(n)} - pT_{ij}^{(n)}}{(n-2)p(2-p)^2} + R_{ij}^{(n)}\right) \\ &= \frac{p}{2-p} + \frac{2}{n(n-1)(n-2)(2-p)^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k\ne i,j}\left(\frac{2\mathbf{1}_{ik}\mathbf{1}_{jk}}{p} - (\mathbf{1}_{ik} + \mathbf{1}_{jk})\right) + \frac{2}{n(n-1)}R_n \\ &= \frac{p}{2-p} - \frac{4P_{1,n}}{n(n-1)(2-p)^2} + \frac{4P_{2,n}}{n(n-1)(n-2)p(2-p)^2} + \frac{2}{n(n-1)}R_n, \end{aligned} \tag{19}$$

where

$$P_{1,n} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{1}_{ij}, \qquad P_{2,n} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k\ne i,j}\mathbf{1}_{ij}\mathbf{1}_{ik}$$

denote the number of edges and the number of paths of length two in $G(n, p)$, respectively. Further, we can rewrite (19) as

$$\frac{(n-1)(2-p)^2}{\sqrt{8p(1-p)}}\left(J_n - \frac{p}{2-p}\right) = \sqrt{\frac{2p}{1-p}}\left(-\frac{P_{1,n}}{np} + \frac{P_{2,n}}{n(n-2)p^2}\right) + \frac{(2-p)^2}{n\sqrt{2p(1-p)}}R_n. \tag{20}$$

For $P_{1,n}$ and $P_{2,n}$, it is not hard to obtain that their expectations are given by $\mathbb{E}[P_{1,n}] = \frac{1}{2}n(n-1)p$ and $\mathbb{E}[P_{2,n}] = \frac{1}{2}n(n-1)(n-2)p^2$. Applying [10, Theorem 3(iii)] yields that if $np \to \infty$ and $n^2(1-p) \to \infty$,

$$\sqrt{\frac{2p}{1-p}}\left(\frac{P_{1,n} - \frac{1}{2}n(n-1)p}{np}, \frac{P_{2,n} - \frac{1}{2}n(n-1)(n-2)p^2}{2n(n-2)p^2}\right) \xrightarrow{\mathrm{D}} (Z, Z),$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

which implies that

$$\sqrt{\frac{2p}{1-p}}\left(-\frac{P_{1,n}}{np}+\frac{P_{2,n}}{n(n-2)p^2}\right)\xrightarrow{\mathrm{D}} Z. \tag{21}$$

To prove Theorem 2, by (20) and (21) it is sufficient to show that

$$\frac{R_n}{n\sqrt{p(1-p)}}\xrightarrow{\mathrm{P}} 0.$$

That is, by Chebyshev's inequality and the fact that $\mathbb{E}[R_n]=0$, we only need to prove that

$$\mathrm{Var}[R_n]=o\big(n^2p(1-p)\big). \tag{22}$$

On the other hand, by symmetry, it follows by (18) that

$$\begin{aligned}\mathrm{Var}[R_n]&=\mathrm{cov}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}R_{ij}^{(n)},\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}R_{ij}^{(n)}\right)\\&=\frac{n(n-1)}{2}\mathrm{cov}\left(R_{12}^{(n)},\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}R_{ij}^{(n)}\right)\\&=\frac{1}{2}n(n-1)\mathrm{Var}\left[R_{12}^{(n)}\right]+n(n-1)(n-2)\mathrm{cov}\left(R_{12}^{(n)},R_{13}^{(n)}\right)\\&\quad+\frac{1}{4}n(n-1)(n-2)(n-3)\mathrm{cov}\left(R_{12}^{(n)},R_{34}^{(n)}\right).\end{aligned} \tag{23}$$

To prove (22), we next estimate the variance and covariances in (23) separately. By the law of total expectation and (17), for the variance of $R_{12}^{(n)}$ we have

$$\mathrm{Var}\left[R_{12}^{(n)}\right]=\mathbb{E}\left[\left(R_{12}^{(n)}\right)^2\right]=\mathbb{E}\left[\mathbb{E}\left[\left(R_{12}^{(n)}\right)^2\mid T_{12}^{(n)}\right]\right]=\sum_{m=0}^{n-2}\mathbb{E}\left[\left(R_{12}^{(n)}\right)^2\mid T_{12}^{(n)}=m\right]\mathbb{P}\left(T_{12}^{(n)}=m\right). \tag{24}$$

Recalling that $T_{12}^{(n)}\sim\mathrm{Bin}(n-2,p(2-p))$, and $R_{12}^{(n)}=0$ if $T_{12}^{(n)}=0$, By (3), (4), and (16), it follows that

$$\begin{aligned}\mathbb{E}\left[\left(R_{12}^{(n)}\right)^2\mid T_{12}^{(n)}=m\right]&=\frac{1}{(n-2)^2p^2(2-p)^2}\left(\frac{(n-2)p(2-p)}{m}-1\right)^2\mathrm{Var}\left(S_{12}^{(n)}\mid T_{12}^{(n)}=m\right)\\&=\frac{2(1-p)}{(n-2)^2p(2-p)^4}\left(\frac{(n-2)^2p^2(2-p)^2}{m}-2(n-2)p(2-p)+m\right),\end{aligned}$$

which, together with (6) and (24), implies that

$$\begin{aligned}\mathrm{Var}\left[R_{12}^{(n)}\right]&=\frac{2(1-p)}{(n-2)^2p(2-p)^4}\sum_{m=1}^{n-2}\left(\frac{(n-2)^2p^2(2-p)^2}{m}-2(n-2)p(2-p)+m\right)\mathbb{P}\left(T_{12}^{(n)}=m\right)\\&=\frac{2(1-p)}{(n-2)(2-p)^3}\left((n-2)p(2-p)\sum_{m=1}^{n-2}\frac{1}{m}\mathbb{P}\left(T_{12}^{(n)}=m\right)-1+2\mathbb{P}\left(T_{12}^{(n)}=0\right)\right)\\&=O\left(\frac{1-p}{n^2p}\right),\end{aligned} \tag{25}$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

where in the last equality we used the simple fact that

$$0 < \mathbb{P}\Big(T_{12}^{(n)} = 0\Big) = (1-p)^{2(n-2)} \le \mathrm{e}^{-2(n-2)p} = o\bigg(\frac{1}{np}\bigg).$$

To calculate the covariance of $R_{12}^{(n)}$ and $R_{13}^{(n)}$, by convention we introduce the shorthand notation

$$\widetilde{T}_{ij}^{(n)} := \frac{(n-2)p(2-p)}{T_{ij}^{(n)}} - 1$$

for distinct vertices $i, j \in [n]$. Recalling (8), we have

$$(2-p)S_{ij}^{(n)} - pT_{ij}^{(n)} = \sum_{k \neq i,j} V_{ij,k}, \quad i, j \in [n].$$

By symmetry and (16), it thus follows that

$$\begin{aligned}
\operatorname{cov}\Big(R_{12}^{(n)}, R_{13}^{(n)}\Big) &= \frac{1}{(n-2)^2p^2(2-p)^4} \sum_{k=3}^{n} \sum_{l \neq 1,3} \mathbb{E}\Big[V_{12,k}V_{13,l}\widetilde{T}_{12}^{(n)}\widetilde{T}_{13}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{13}^{(n)} > 0\Big)\Big] \\
&= \frac{1}{(n-2)p^2(2-p)^4}\mathbb{E}\Big[V_{12,3}V_{13,2}\widetilde{T}_{12}^{(n)}\widetilde{T}_{13}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{13}^{(n)} > 0\Big)\Big] \\
&\quad + \frac{(n-3)}{(n-2)p^2(2-p)^4}\mathbb{E}\Big[V_{12,3}V_{13,4}\widetilde{T}_{12}^{(n)}\widetilde{T}_{13}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{13}^{(n)} > 0\Big)\Big], \qquad (26)
\end{aligned}$$

which splits the covariance into two parts. Notice that, by (8), the discrete random variable $V_{ij,k} \neq 0$ if and only if $\mathbf{1}_{ik} \vee \mathbf{1}_{jk} = 1$. This implies that if $V_{12,3}V_{13,2} \neq 0$, we have $T_{12}^{(n)} = 1 + \sum_{k=4}^{n} \mathbf{1}_{1k} \vee \mathbf{1}_{2k}$ and $T_{13}^{(n)} = 1 + \sum_{k=4}^{n} \mathbf{1}_{1k} \vee \mathbf{1}_{3k}$, and they have the same distribution. Since it follows that, by conditioning on $\mathbf{1}_{23}$,

$$\begin{aligned}
\mathbb{E}[V_{12,3}V_{13,2}] &= \mathbb{E}\big[\big((2-p)\mathbf{1}_{13}\mathbf{1}_{23} - p(\mathbf{1}_{13} \vee \mathbf{1}_{23})\big)\big((2-p)\mathbf{1}_{12}\mathbf{1}_{23} - p(\mathbf{1}_{12} \vee \mathbf{1}_{23})\big)\big] \\
&= p\mathbb{E}\big[\big((2-p)\mathbf{1}_{13} - p\big)\big((2-p)\mathbf{1}_{12} - p\big)\big] + (1-p)\mathbb{E}\big[p^2\mathbf{1}_{13}\mathbf{1}_{12}\big] \\
&= p^3(1-p)^2 + p^4(1-p) = p^3(1-p),
\end{aligned}$$

and that, by Lemma 1 and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\bigg|\mathbb{E}\bigg[\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{2k}} - 1\bigg)\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{3k}} - 1\bigg)\bigg]\bigg| \\
&\qquad\qquad \le \mathbb{E}\bigg[\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{2k}} - 1\bigg)^2\bigg] = O\bigg(\frac{1}{np}\bigg),
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}\Big[V_{12,3}V_{13,2}\widetilde{T}_{12}^{(n)}\widetilde{T}_{13}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{13}^{(n)} > 0\Big)\Big] \\
&\quad = \mathbb{E}\bigg[V_{12,3}V_{13,2}\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{2k}} - 1\bigg)\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{3k}} - 1\bigg)\bigg] \\
&\quad = p^3(1-p)\mathbb{E}\bigg[\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{2k}} - 1\bigg)\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k=4}^{n}\mathbf{1}_{1k}\vee\mathbf{1}_{3k}} - 1\bigg)\bigg] \\
&\quad = O\bigg(\frac{p^2(1-p)}{n}\bigg). \qquad (27)
\end{aligned}$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

Let us define $\mathcal{B}_{ab} := \{\mathbf{1}_{12} \vee \mathbf{1}_{23} = a,\ \mathbf{1}_{14} \vee \mathbf{1}_{24} = b\}$, $a, b \in \{0, 1\}$. Noting that $\mathbb{E}[V_{12,3}] = 0$, we have

$$\mathbb{E}\big[V_{12,3}\mathbf{1}(\mathbf{1}_{12} \vee \mathbf{1}_{23} = 1)\big] = -\mathbb{E}\big[V_{12,3}\mathbf{1}(\mathbf{1}_{12} \vee \mathbf{1}_{23} = 0)\big] = p\mathbb{E}\big[\mathbf{1}_{13}\mathbf{1}(\mathbf{1}_{12} = \mathbf{1}_{23} = 0)\big] = p^2(1-p)^2,$$

which implies that, for any $a, b = 0$ or 1,

$$\begin{aligned}\mathbb{E}[V_{12,3}V_{13,4}\mathbf{1}(\mathcal{B}_{ab})] &= \mathbb{E}[V_{12,3}\mathbf{1}(\mathbf{1}_{12} \vee \mathbf{1}_{23} = a)]\mathbb{E}[V_{13,4}\mathbf{1}(\mathbf{1}_{14} \vee \mathbf{1}_{24} = b)] \\ &= \mathbb{E}[V_{12,3}\mathbf{1}(\mathbf{1}_{12} \vee \mathbf{1}_{23} = a)]\mathbb{E}[V_{12,3}\mathbf{1}(\mathbf{1}_{12} \vee \mathbf{1}_{23} = b)] \\ &= (-1)^{a+b}p^4(1-p)^4.\end{aligned}$$

Analogously to (27), we have

$$\begin{aligned}&\mathbb{E}\Big[V_{12,3}V_{13,4}\widetilde{T}_{12}^{(n)}\widetilde{T}_{13}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{13}^{(n)} > 0\Big)\Big] \\ &\quad= \sum_{a=0}^{1}\sum_{b=0}^{1}\mathbb{E}\Big[V_{12,3}V_{13,4}\widetilde{T}_{12}^{(n)}\widetilde{T}_{13}^{(n)}\mathbf{1}(\mathcal{B}_{ab})\Big] \\ &\quad= \sum_{a=0}^{1}\sum_{b=0}^{1}\mathbb{E}\big[V_{12,3}V_{13,4}W_{12}(b)W_{13}(a)\mathbf{1}(\mathcal{B}_{ab})\big] \\ &\quad= \sum_{a=0}^{1}\sum_{b=0}^{1}\mathbb{E}\big[V_{12,3}V_{13,4}\mathbf{1}(\mathcal{B}_{ab})\big]\mathbb{E}\big[W_{12}(b)W_{13}(a)\big] \\ &\quad= p^4(1-p)^4\mathbb{E}\big[(W_{12}(0) - W_{12}(1))(W_{13}(0) - W_{13}(1))\big],\end{aligned} \tag{28}$$

where

$$W_{ij}(a) := \frac{(n-2)p(2-p)}{1 + a + \sum_{k=5}^{n}\mathbf{1}_{ik} \vee \mathbf{1}_{jk}} - 1, \quad i, j \in [n],\ a = 0, 1.$$

Noting that $W_{12}(0) - W_{12}(1)$ and $W_{13}(0) - W_{13}(1)$ have the same distribution, by Lemma 2 and the Cauchy–Schwarz inequality we have

$$\begin{aligned}\mathbb{E}\big|\big(W_{12}(0) - W_{12}(1)\big)\big(W_{13}(0) - W_{13}(1)\big)\big| &\le \mathbb{E}\big[(W_{12}(0) - W_{12}(1))^2\big] \\ &\le (n-2)^2p^2(2-p)^2\mathbb{E}\left[\frac{1}{\Big(1 + \sum_{k=5}^{n}\mathbf{1}_{1k} \vee \mathbf{1}_{2k}\Big)^4}\right] \\ &= O\Big(\frac{1}{n^2p^2}\Big),\end{aligned}$$

which, together with (26), (27), and (28), implies that

$$\operatorname{cov}\Big(R_{12}^{(n)}, R_{13}^{(n)}\Big) = O\Big(\frac{1-p}{n^2}\Big). \tag{29}$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

It remains to calculate the second covariance $\mathbf{Cov}\Big(R_{12}^{(n)}, R_{34}^{(n)}\Big)$ on the right-hand side of (23), and the procedure is similar to the previous one. We sketch the calculations below, omitting a few specific interpretations. Analogously to (26), we have

$$\begin{aligned}\operatorname{cov}\Big(R_{12}^{(n)}, R_{34}^{(n)}\Big) &= \frac{1}{(n-2)^2p^2(2-p)^4}\sum_{k=3}^{n}\sum_{l\neq 3,4}\mathbb{E}\Big[V_{12,k}V_{34,l}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{34}^{(n)}>0\Big)\Big]\\ &= \frac{4}{(n-2)^2p^2(2-p)^4}\mathbb{E}\Big[V_{12,3}V_{34,1}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{34}^{(n)}>0\Big)\Big]\\ &\quad+\frac{2}{(n-2)p^2(2-p)^4}\mathbb{E}\Big[V_{12,3}V_{34,5}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{34}^{(n)}>0\Big)\Big]\\ &\quad+\frac{(n-4)^2}{(n-2)^2p^2(2-p)^4}\mathbb{E}\Big[V_{12,5}V_{34,5}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{34}^{(n)}>0\Big)\Big]. \end{aligned}\tag{30}$$

Define $\mathcal{C}_{ab} := \{\mathbf{1}_{14}\vee\mathbf{1}_{24}=a,\ \mathbf{1}_{23}\vee\mathbf{1}_{24}=b\}$, $a, b\in\{0,1\}$. After straightforward calculations, we have

$$\begin{aligned}&\mathbb{E}\big[V_{12,3}V_{34,1}\mathbf{1}(\mathcal{C}_{00})\big]=p^3(1-p)^3,\\ &\mathbb{E}\big[V_{12,3}V_{34,1}\mathbf{1}(\mathcal{C}_{11})\big]=p^3(1-p)\big(1+3(1-p)^2\big),\\ &\mathbb{E}\big[V_{12,3}V_{34,1}\mathbf{1}(\mathcal{C}_{01})\big]=\mathbb{E}\big[V_{12,3}V_{34,1}\mathbf{1}(\mathcal{C}_{10})\big]=-2p^3(1-p)^3.\end{aligned}$$

Then we can conclude that $\mathbb{E}[V_{12,3}V_{34,1}\mathbf{1}(\mathcal{C}_{ab})]=O(p^3(1-p))$. Hence, by the Cauchy–Schwarz inequality and Lemma 1,

$$\begin{aligned}\mathbb{E}\Big[V_{12,3}V_{34,1}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{34}^{(n)}>0\Big)\Big] &= \sum_{a=0}^{1}\sum_{b=0}^{1}\mathbb{E}\big[V_{12,3}V_{34,1}W_{12}(a)W_{34}(b)\mathbf{1}(\mathcal{C}_{ab})\big]\\ &= \sum_{a=0}^{1}\sum_{b=0}^{1}\mathbb{E}[V_{12,3}V_{34,1}\mathbf{1}(\mathcal{C}_{ab})]\mathbb{E}[W_{12}(a)]\mathbb{E}[W_{34}(b)]\\ &= O\bigg(\frac{p^2(1-p)}{n}\bigg).\end{aligned}\tag{31}$$

Since $V_{12,3}$ and $V_{34,5}$ are independent and have the common mean 0, it follows that

$$\begin{aligned}&\mathbb{E}\Big[V_{12,3}V_{34,5}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\Big(T_{12}^{(n)}T_{34}^{(n)}>0\Big)\Big]\\ &=\mathbb{E}\Big[V_{12,3}V_{34,5}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}(\mathbf{1}_{13}\vee\mathbf{1}_{23}=1)\mathbf{1}(\mathbf{1}_{35}\vee\mathbf{1}_{45}=1)\Big]\\ &=\mathbb{E}[V_{34,5}\mathbf{1}(\mathbf{1}_{35}\vee\mathbf{1}_{45}=1)]\mathbb{E}\bigg[V_{12,3}\widetilde{T}_{12}^{(n)}\mathbf{1}(\mathbf{1}_{13}\vee\mathbf{1}_{23}=1)\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k\neq3,4,5}(\mathbf{1}_{3k}\vee\mathbf{1}_{4k})}-1\bigg)\bigg]\\ &=\mathbb{E}[V_{34,5}]\mathbb{E}\bigg[V_{12,3}\widetilde{T}_{12}^{(n)}\mathbf{1}(\mathbf{1}_{13}\vee\mathbf{1}_{23}=1)\bigg(\frac{(n-2)p(2-p)}{1+\sum_{k\neq3,4,5}(\mathbf{1}_{3k}\vee\mathbf{1}_{4k})}-1\bigg)\bigg]=0.\end{aligned}\tag{32}$$

Similarly, we also have $\mathbb{E}\big[V_{12,5}V_{34,5}\widetilde{T}_{12}^{(n)}\widetilde{T}_{34}^{(n)}\mathbf{1}\big(T_{12}^{(n)}T_{34}^{(n)}>0\big)\big]=0$. Plugging this, (31), and (32) into (30) yields

$$\operatorname{cov}\Big(R_{12}^{(n)}, R_{34}^{(n)}\Big)=O\bigg(\frac{1-p}{n^3}\bigg),$$

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

which, together with (23), (25), and (29), implies that

$$\mathrm{Var}[R_n] = O\left(\frac{1-p}{p}\right) + O(n(1-p)) + O(n(1-p)) = O(n(1-p))$$

as $np \to \infty$. This proves (22), and thus completes the proof of Theorem 2. □

Acknowledgements

We wish to thank two anonymous referees for their constructive comments that helped improve the quality of the paper.

Funding information

This work is supported by NSFC grant numbers 11671373 and 11771418.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] ALI, M. *et al.* (2021). Machine learning – a novel approach of well logs similarity based on synchronization measures to predict shear sonic logs. *J. Petroleum Sci, Eng.* **203**, 108602.

[2] ARIAS-CASTRO, E. AND VERZELEN, N. (2014). Community detection in dense random networks. *Ann. Statist.* **42**, 940–969.

[3] BAG, S., KUMAR, S. K. AND TIWARI, M. K. (2019). An efficient recommendation generation using relevant Jaccard similarity. *Inf. Sci.* **483**, 53–64.

[4] BERAHMAND, K., BOUYER, A. AND VASIGHI, M. (2018). Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes. *IEEE Trans. Comput. Soc. Syst.* **5**, 1021–1033.

[5] BOLLOBÁS, B. (2001). *Random Graphs*, 2nd edn. Cambridge University Press.

[6] CHUNG, N. C., MIASOJEDOW, B., STARTEK, M. AND GAMBIN, A. (2019). Jaccard/Tanimoto similarity test and estimation methods for biological presence–absence data. *BMC Bioinform.* **20**, 1–11.

[7] DA FONTOURA COSTA, L. (2021). Further generalizations of the Jaccard index. Preprint, arXiv:2110.09619.

[8] EELBODE, T. *et al.* (2020). Optimization for medical image segmentation: Theory and practice when evaluating with dice score or Jaccard index. *IEEE Trans. Med. Imag.* **39**, 3679–3690.

[9] FAN, X. *et al.* (2019). Similarity and heterogeneity of price dynamics across China's regional carbon markets: A visibility graph network approach. *Appl. Energy* **235**, 739–746.

[10] FENG, Q., HU, Z. AND SU, C. (2013). The Zagreb indices of random graphs. *Prob. Eng. Inf. Sci.* **27**, 247–260.

[11] GILBERT, G. (1972). Distance between sets. *Nature* **239**, 174.

[12] HENNIG, C. (2007). Cluster-wise assessment of cluster stability. *Comput. Statist. Data Anal.* **52**, 258–271.

[13] JACCARD, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* **37**, 241–272.

[14] JANSON, S., LUCZAK, T. AND RUCINSKI, A. (2000). *Random Graphs*. John Wiley, New York.

[15] KOENEMAN, S. H. AND CAVANAUGH, J. E. (2022). An improved asymptotic test for the Jaccard similarity index for binary data. *Statist. Prob. Lett.* **184**, 109375.

[16] KOGGE, P. M. (2016). Jaccard coefficients as a potential graph benchmark. In *Proc. 2016 IEEE Int. Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 921–928. IEEE, Piscataway, NJ.

[17] KOSUB, S. (2019). A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Lett.* **120**, 36–38.

[18] LU, H. AND UDDIN, S. (2023). Embedding-based link predictions to explore latent comorbidity of chronic diseases. *Health Inf. Sci. Syst.* **11**, 2.

[19] MAMMONE, N. *et al.* (2018). Permutation Jaccard distance-based hierarchical clustering to estimate EEG network density modifications in MCI subjects. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 5122–5135.

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press

[20] MIASNIKOF, P., SHESTOPALOFF, A. Y., PITSOULIS, L. AND PONOMARENKO, A. (2022). An empirical comparison of connectivity-based distances on a graph and their computational scalability. *J. Complex Netw.* **10**, cnac003.
[21] SATHRE, P., GONDHALEKAR, A. AND FENG, W. C. (2022). Edge-connected Jaccard similarity for graph link prediction on FPGA. In *Proc. 2022 IEEE High Performance Extreme Computing Conf. (HPEC)*, pp. 1–10. IEEE, Piscataway, NY.
[22] SHESTOPALOFF, P. M., ALEXANDER,Y., BRAVO, C. AND LAWRYSHYN, Y. (2023). Statistical network isomorphism. In *Complex Networks and Their Applications XI*, eds H. Cherifi, R. N. Mantegna, L. M. Rocha, C. Cherifi, and S. Micciche. Springer, New York, pp. 325–336.
[23] SHI, X., WU, Y. AND LIU, Y. (2010). A note on asymptotic approximations of inverse moments of nonnegative random variables. *Statist. Prob. Lett.* **80**, 1260–1264.
[24] SINGH, M. D., KRISHNA, P. R. AND SAXENA, A. (2009). A privacy preserving Jaccard similarity function for mining encrypted data. In *Proc. TENCON 2009–2009 IEEE Region 10 Conf.*, pp. 1–4.
[25] VAN DER HOFSTAD, R. (2016). *Random Graphs and Complex Networks*. Cambridge University Press.
[26] VERZELEN, N. AND ARIAS-CASTRO, E. (2015). Community detection in sparse random networks. *Ann. Appl. Prob.* **25**, 3465–3510.
[27] WU, C. AND WANG, B. (2017). Extracting topics based on word2vec and improved Jaccard similarity coefficient. In *Proc. 2017 IEEE 2nd Int. Conf. Data Science in Cyberspace (DSC)*, pp. 389–397.
[28] Wuyungaowa and Wang, T. (2008). Asymptotic expansions for inverse moments of binomial and negative binomial. *Statist. Prob. Lett.* **78**, 3018–3022.
[29] YIN, Y. AND YASUDA, K. (2006). Similarity coefficient methods applied to the cell formation problem: A taxonomy and review. *Int. J. Prod. Econ.* **101**, 329–352.
[30] ZHANG, P. *et al.* (2016). Measuring the robustness of link prediction algorithms under noisy environment. *Sci. Rep.* **6**, 18881.

https://doi.org/10.1017/jpr.2023.112 Published online by Cambridge University Press