

Breakpoint medians and anti-medians for signed genomes

Poly H da Silva¹, Arash Jamshidpey², and David Sankoff³

¹ Independent researcher, USA
polyhdasilva@gmail.com

² Department of Statistics, University of California at Berkeley, CA, USA
arash.jamshidpey@berkeley.edu

³ Department of Mathematics and Statistics, University of Ottawa, ON, Canada
sankoff@uottawa.ca

Abstract. The breakpoint distance is a fundamental measure for comparing gene orders and plays a central role in genome rearrangement phylogenetics. For more than two genomes, breakpoint medians are widely used to infer ancestral gene orders, yet their typical behavior is difficult to characterize due to the complex, non-geodesic geometry of genome space. A conjecture of Haghghi and Sankoff states that for random genomes, breakpoint medians tend to lie near one of the input genomes rather than constituting genuinely intermediate solutions. In this paper, we study breakpoint medians of independently and uniformly sampled signed multichromosomal genomes. We prove the Haghghi–Sankoff conjecture in this setting for the first time. We show that with high probability, any breakpoint median must remain close to a single input genome, rather than drawing adjacencies from multiple genomes. Our analysis relies on probabilistic bounds for usable adjacencies in random signed genomes and structural constraints of the signed multichromosomal model. We further provide a game-theoretic interpretation that explains why compromise strategies lead to increased total distance and the emergence of anti-medians.

Keywords: Genomes · Breakpoint distance · Medians · Anti-medians.

1 Introduction

Breakpoint distance, introduced in [10] as a combinatorial measure of gene-order dissimilarity, has proved to be a powerful tool for studying similarities between genomes. For more than two genomes, the notion of a median genome plays a central role in comparative genomics, particularly in the reconstruction of gene-order phylogenies. In the breakpoint framework, the median problem was introduced by Sankoff *et al.* [11], who showed that computing a breakpoint median reduces to the Traveling Salesman Problem (TSP) making it NP-hard for unichromosomal unsigned genomes. Since then, the breakpoint median problem has been extensively studied [1, 2, 15, 5, 4, 16, 18].

The difficulty of finding breakpoint medians is rooted in the complex geometry of genome space endowed with the breakpoint distance. More precisely, the breakpoint distance is not geodesic, meaning that midpoints between genomes may fail to exist. As a result, shortest paths need not behave continuously, and standard geometric intuition breaks down (see [7] for an accessible discussion). This lack of geodesic structure significantly complicates the median problem: algorithms that attempt to approach a median iteratively, starting from an input genome, often fail. Although many heuristic algorithms for finding breakpoint medians have been proposed, they are insufficient for exploring the global geometry of genome space or for understanding the typical positions and density of breakpoint medians of sampled genomes.

Using heuristic algorithms, Haghghi and Sankoff [6] observed an unexpected phenomenon: for k randomly sampled unichromosomal unsigned genomes with n genes, the breakpoint median value is approximately $(k - 1)n$. Even more surprisingly, their simulation studies suggested that medians tend to lie near the “corners” of genome space, meaning that most heuristic medians were positioned very close to one of the input genomes. Nevertheless, a small proportion of medians appeared far from the corners, with some even roughly equidistant from all inputs. Identifying and understanding such “midpoint” medians is of particular interest, as these genomes potentially carry balanced genomic information from all inputs and may be relevant for ancestral genome reconstruction.

In subsequent work, Jamshidpey *et al.* [7] partially proved the Haghghi–Sankoff conjecture by establishing the median value $(k - 1)n$. That work also introduced the notions of partial geodesics (or geodesic patches) as a means of reaching non-trivial medians far from the corners, termed “accessible genomes”. A key result was that almost all adjacencies of a median must already appear among the adjacencies of at least one input genome. In other words, medians are not allowed to introduce many new adjacencies outside the union of adjacencies of the input genomes. This observation is crucial: it opens a principled pathway to constructing medians far from corners by selecting adjacencies from all input genomes, rather than predominantly from a single one.

Motivated by this observation, Larlee *et al.* [9, 8] attempted to construct medians by compromising the number of adjacencies taken from k randomly sampled signed genomes. Specifically, for $x \in (0, 1)$, they sought to construct a genome G that takes a proportion x of adjacencies from each of the k input genomes, yielding a total of kx adjacencies, and to maximize x subject to the constraint that kx be as close as possible to 1. Unexpectedly, this approach failed and produced the opposite effect. Larlee *et al.* [9, 8] observed that the more one attempts to compromise, the smaller the maximal achievable value of x becomes, and hence the greater the total distance of G from the input genomes. They introduced the notion of anti-medians to describe genomes that are equidistant from all inputs yet as far as possible from being medians.

Random sampling of adjacencies from input genomes [8] offers little control over the structural compatibility of adjacencies in signed multichromosomal genomes with circular chromosomes. The original Haghghi–Sankoff conjecture,

however, concerned unichromosomal unsigned linear genomes, and additional constraints arising from unichromosomal structure must be handled carefully. Motivated by the work of Larlee *et al.* [8], da Silva *et al.* [12] introduced necessary conditions and constraints governing how adjacencies can be taken from each input genome. We believe that the framework developed in [12] will play a key role in answering whether the Haghghi–Sankoff conjecture holds or fails for unichromosomal unsigned genomes. By contrast, the case of multichromosomal signed genomes is more tractable, and in this setting, we are able to fully prove the Haghghi–Sankoff conjecture in this work.

As discussed above, understanding how far medians can lie from the corners is deeply tied to the geometry of genome space. Moreover, the technical tools required for signed versus unsigned, and for unichromosomal versus multichromosomal genomes, differ substantially. In particular, the constraints involved in constructing signed genomes, which are the focus of this paper, are significantly stronger than those arising in the unsigned setting. This added structure makes it possible to prove versions of the Haghghi–Sankoff conjecture for signed genomes that remain out of reach for unsigned ones at this moment. Nonetheless, the large body of heuristic algorithms for breakpoint medians (many based on TSP heuristics) remains insufficient for understanding the geometry of the breakpoint distance and the relative position of medians relative to input genomes.

A recent breakthrough in this direction was achieved in our earlier work [13], where for the first time all exact (non-heuristic) medians of randomly sampled unichromosomal genomes were enumerated (the method applies to multichromosomal genomes as well). This exhaustive enumeration enabled a much more precise understanding of the geometry of genome space and of the positions of medians relative to the sampled input genomes.

In this paper, we study the relative position of breakpoint medians for k independently and uniformly sampled signed multichromosomal genomes. For simplicity, we assume that all chromosomes are circular; the methods extend naturally to linear chromosomes via a standard telomere-capping construction. For the first time, we prove the Haghghi–Sankoff conjecture for signed multichromosomal genomes. More precisely, we show that with high probability, for any median genome \mathcal{G} of k independently and randomly sampled genomes $\mathcal{G}_1, \dots, \mathcal{G}_k$, if x_i denotes the proportion of adjacencies shared between \mathcal{G} and \mathcal{G}_i , with $x_1 + \dots + x_k = 1$, then there exists an index j such that $x_j = 1$ and $x_i = 0$ for all $i \neq j$. Equivalently, any median must lie very close to one of the input genomes. Our results continue the line of work initiated by [9, 8], but differ fundamentally from attempts to prove analogous statements for unsigned unichromosomal genomes, such as the approach in [12]. The stricter constraints in signed multichromosomal genome models play a decisive role in enabling our proofs.

More precisely, motivated by similar ideas for signed genomes [8], da Silva *et al.* [12] introduced a classification of gene adjacencies of a random genome \mathcal{G}' relative to a given reference set of adjacencies \mathcal{I} . This classification allows one to count the number of usable adjacencies of \mathcal{G}' with respect to \mathcal{I} .

More generally, in the context of constructing a median genome for k independently and randomly sampled genomes $\mathcal{G}_1, \dots, \mathcal{G}_k$, this implies that any median \mathcal{G} must draw almost all of its adjacencies from at least one of the input genomes. Suppose that a set of adjacencies \mathcal{I}_r has already been taken from $\mathcal{G}_1, \dots, \mathcal{G}_r$, for $1 \leq r < k$. One then seeks the largest possible set of usable adjacencies \mathcal{J}_{r+1} from \mathcal{G}_{r+1} such that $\mathcal{I}_{r+1} = \mathcal{I}_r \cup \mathcal{J}_{r+1}$ can still be extended to a median genome. The goal is to optimize this sampling process so that \mathcal{I}_k constitutes a complete genome \mathcal{G} . Finding such a \mathcal{G} guarantees that it is a median of $\mathcal{G}_1, \dots, \mathcal{G}_k$.

This is not always possible, and depending on the underlying restrictions, the size difference between a full genome and \mathcal{I}_k may be substantial. These restrictions depend on the genome model, including whether genomes are signed or unsigned, and multichromosomal or unichromosomal. Comparing signed and unsigned cases, once a set of adjacencies \mathcal{I} has been sampled, the average number of usable adjacencies of a random genome is significantly smaller in the signed case. This observation is the key mechanism that allows us to prove the Haghghi–Sankoff conjecture for signed multichromosomal genomes in this paper.

Finally, we use game-theoretic ideas to explain why compromising, as in [8], leads to increased distance rather than improved median quality. We model each genome as a player in a cooperative general-sum game with transferable utility. Each player \mathcal{G}_i gains a payoff x_i , the proportion of its adjacencies used in the constructed genome \mathcal{G} , and the players jointly aim to maximize $x_1 + \dots + x_k \leq 1$. We determine the Pareto-optimal boundary of this game and show that the maximum total payoff is achieved at the extreme points $e_i = (0, \dots, 1, \dots, 0)$. As one moves along the Pareto boundary toward more balanced compromises, the total payoff decreases, reaching its minimum at a symmetric point $x = (a, \dots, a)$ with $a < 1$. This fully explains why increased compromise in the number of gene adjacencies taken from input genomes leads to genomes that are farther from being medians and clarifies the emergence of anti-medians.

The remainder of the paper is organized as follows. In Section 2, we review genome representations, breakpoint distance for signed genomes, and the behavior of breakpoint distance for two random genomes. In this section, we also establish probabilistic results for medians of k independently sampled genomes. In Section 3, we prove the Haghghi–Sankoff conjecture in the signed multichromosomal setting and show that, with high probability, all approximate (and hence exact) medians lie at the corners. Finally, in Section 4, we develop the game-theoretic framework explaining why compromise in the number of gene adjacencies taken from input genomes increases the total distance and leads to the appearance of anti-medians.

2 Probabilistic structure of common adjacencies and breakpoint medians

Although the methods in this paper can be easily extended to multichromosomal genomes with linear and circular chromosomes, for the sake of simplicity we assume that all chromosomes are circular. Therefore, having no telom-

ere in our model, a signed multichromosomal genome with circular chromosomes can be represented as a perfect matching on the set of vertices $\pm[n] := \{\pm 1, \pm 2, \dots, \pm n\}$.

More precisely, for a set of genes labeled by $1, 2, \dots, n$, we denote by $-i$ and $+i$ the two extremities of gene i . In fact the signs may be interpreted as encoding gene polarity (strandedness). A genome with gene content $1, 2, \dots, n$ is then represented by a perfect matching on $\pm[n]$, where each edge of the matching determines a gene adjacency, linking two neighbouring genes. Note that allowing chromosomes in arbitrary number and of arbitrary sizes places no restrictions on the associated perfect matching. For instance, an edge (gene adjacency) $\{-i, +i\}$ means that the genome contains a circular chromosome of size 1, with gene content i . We denote by \mathbb{G}_n the set of all perfect matchings on $\pm[n]$.

In general, the gene content of a genome G may be any finite subset $V(G) \subseteq [n]$. In this case, the gene extremities of G are elements of $\pm V(G) := V(G) \cup (-V(G))$, and the genome G is a perfect matching on $\pm V(G)$. We represent such a genome as $G = (\pm V(G), E(G))$, where $\pm V(G)$ is the set of vertices and $E(G)$ is its set of edges. Note that for $G \in \mathbb{G}_n$, $\pm V(G) = \pm[n]$.

Introduced in [10], breakpoint distance (bp-distance) is often used to measure the similarity between two genomes. More specifically, for a pair of genomes G, G' , the breakpoint distance $d(G, G')$ can be defined by

$$d(G, G') = \frac{1}{2} |E(G) \triangle E(G')|,$$

where \triangle denotes the symmetric difference between sets and $|\cdot|$ denotes the cardinality of the set. In particular, for $G, G' \in \mathbb{G}_n$, this formula reduces to

$$d(G, G') = n - |E(G) \cap E(G')|.$$

Letting \mathbb{G} be the space of all perfect matchings on $\pm S := S \cup (-S)$, for finite subsets $S \subset \mathbb{N}$, (\mathbb{G}, d) is a metric space. In addition, for any $n \in \mathbb{N}$, (\mathbb{G}_n, d) is also a metric space. Let $\mathbf{S}_{\pm n}$ be the group of permutations on $\pm[n]$ with the function composition as the group action. We can define the action of $\mathbf{S}_{\pm n}$ on \mathbb{G}_n by $\sigma \cdot G \in \mathbb{G}_n$ for $\sigma \in \mathbf{S}_{\pm n}$ and $G \in \mathbb{G}_n$, where

$$\pm V(\sigma \cdot G) = \pm[n], \quad E(\sigma \cdot G) = \{\{\sigma(i), \sigma(j)\} : \{i, j\} \in E(G)\}.$$

That is, the action relabels the gene extremities via the permutation $\sigma \in \mathbf{S}_{\pm n}$.

Indeed, the action of $\mathbf{S}_{\pm n}$ on \mathbb{G}_n defined above is distance preserving, i.e., for any $G, G' \in \mathbb{G}_n$ and $\sigma \in \mathbf{S}_{\pm n}$, $d(\sigma \cdot G, \sigma \cdot G') = d(G, G')$. We say that the bp-distance is invariant under the action of $\mathbf{S}_{\pm n}$ on \mathbb{G}_n .

We begin by analyzing the distribution of common adjacencies between independent random genomes. This provides a probabilistic baseline for genome comparison and plays a key role in understanding the structure and typical behavior of breakpoint medians in later sections.

Let \mathcal{G} and \mathcal{G}' be two independent genomes chosen uniformly at random from \mathbb{G}_n . Let $o_n \in \mathbb{G}_n$ be the perfect matching on $\pm[n]$ defined by $E(o_n) = \{\{-i, +i\} :$

$i \in [n]\}$. Let $(\mathcal{G}')^{-1} \subset \mathbf{S}_{\pm n}$ denote the set of all permutations σ on $\pm[n]$ such that $\sigma \cdot \mathcal{G}' = o_n$. Let $\pi_{\mathcal{G}'}$ be selected uniformly at random from $(\mathcal{G}')^{-1}$.

Since \mathcal{G}' is a uniformly random perfect matching in \mathbb{G}_n , the mapping $\pi : \mathbb{G}_n \rightarrow \mathbf{S}_{\pm n}$ defined by $\mathcal{G}' \mapsto \pi_{\mathcal{G}'}$ induces the uniform probability on $\mathbf{S}_{\pm n}$. In other words, $\pi_{\mathcal{G}'}$ is in fact a uniformly random permutation chosen from $\mathbf{S}_{\pm n}$. Furthermore, $\pi_{\mathcal{G}'}$ is independent of \mathcal{G} , and $\pi_{\mathcal{G}'} \cdot \mathcal{G}$ is a uniformly random perfect matching in \mathbb{G}_n . As a result, since d is invariant under the action of $\mathbf{S}_{\pm n}$ on \mathbb{G}_n , we deduce

$$d(\mathcal{G}', \mathcal{G}) \stackrel{d}{=} d(o_n, \pi_{\mathcal{G}'} \cdot \mathcal{G}) \stackrel{d}{=} d(o_n, \mathcal{G}),$$

where “ $\stackrel{d}{=}$ ” indicates equivalence in distribution. This reduction allows us to replace a two-genome comparison with a one-genome problem relative to a fixed reference, greatly simplifying the analysis. More precisely, to study the bp-distance of two independent random genomes, it suffices to study the bp-distance of a uniformly random genome $\mathcal{G} \in \mathbb{G}_n$ and the reference genome $o_n \in \mathbb{G}_n$. Note that the choice of o_n is arbitrary and any other fixed perfect matching in \mathbb{G}_n can serve as the reference genome.

To simplify notation, for the set of genomes $V = \{G_1, \dots, G_m\}$, let

$$\mathcal{A}_V = \mathcal{A}_{G_1, G_2, \dots, G_m} = E(G_1) \cap E(G_2) \cap \dots \cap E(G_m).$$

Intuitively, in large random genomes the probability that a specific adjacency appears in both genomes is small, and distinct adjacencies behave almost independently. It is well-known that for two independent random genomes $\mathcal{G}_n, \mathcal{G}'_n \in \mathbb{G}_n$, as $n \rightarrow \infty$, the number of common adjacencies $|\mathcal{A}_{\mathcal{G}_n, \mathcal{G}'_n}|$ converges in distribution to $\text{Poisson}(1/2)$, i.e. to a Poisson random variable with parameter $1/2$ (see for example [17]). We give a full treatment of this statement in Appendix A, where we show that a stronger mode of convergence—namely, convergence in total variation—holds (cf. Appendix A).

In the sequel, we use the Poisson asymptotics obtained in Appendix A to analyze the typical behavior of breakpoint medians for multiple random genomes. In particular, we study the median of $k \in \mathbb{N}$ perfect matchings $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_k$, picked independently and uniformly at random from \mathbb{G}_n .

The geometric medians were used in comparative genomics for the first time by Sankoff *et al.* [11]. They have extensive applications in gene-order phylogenetics. A breakpoint (bp) median of a set of genomes $A = \{G_1, G_2, \dots, G_k\} \subseteq \mathbb{G}_n$ is a genome $G \in \mathbb{G}_n$ that minimizes the total breakpoint distance function

$$\tilde{G} \mapsto \sum_{i=1}^k d(\tilde{G}, G_i).$$

The minimum value of the function is called the bp-median value of A . We denote by $M_n(A) = M(A) = M(G_1, \dots, G_k)$ the set of all bp-medians of A , and by $\mu_n(A) = \mu(A) = \mu(G_1, \dots, G_k)$ the median value of A . The median of a single genome G is itself. Also, G is a median of two genomes G_1, G_2 if and only if $\mathcal{A}_{G_1, G_2} \subseteq \mathcal{A}_G \subseteq \mathcal{A}_{G_1} \cup \mathcal{A}_{G_2}$ (cf. [7] for unsigned genomes). In addition, for k

genomes $A = \{G_1, G_2, \dots, G_k\} \subseteq \mathbb{G}_n$,

$$\mu(A) \leq \sum_{i=1}^k d(G, G_i) \leq (k-1)n.$$

We continue by recalling that for k perfect matchings $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_k$, chosen independently and uniformly at random from \mathbb{G}_n , the pairwise distances satisfy $d(\mathcal{G}_i, \mathcal{G}_j) = n - O_{\mathbb{P}}(1)$, with Poisson(1/2)-type fluctuations in the number of common adjacencies. Consequently, studying the medians and median value of $\mathcal{G}_1, \dots, \mathcal{G}_k$ is similar to studying those of deterministic genomes G_1, \dots, G_k with pairwise maximum distance $d(G_i, G_j) = n$ for $i \neq j$. The proof of the following proposition is given in Appendix B.

Proposition 1. *For $k, n \geq 1$, let $A = \{G_1, \dots, G_k\} \subset \mathbb{G}_n$ such that $d(G_i, G_j) = n$ for all $i \neq j$. Then $\mu(A) = (k-1)n$, and for any $G \in M(A)$,*

$$\mathcal{A}_G \subseteq \bigcup_{i=1}^k \mathcal{A}_{G_i}.$$

Remark 1. Under the assumption of Proposition 1, $A \subseteq M(A)$.

Proposition 1 shows that when genomes are pairwise at maximum bp-distance, any median is forced to draw all its adjacencies from the input genomes. In particular, no median can introduce new adjacencies beyond those already present in the data. We now extend this deterministic picture to the case of k independent random genomes. Since random genomes are pairwise almost maximally distant, Proposition 1 suggests that their medians should behave similarly, up to small random fluctuations. We make this intuition precise by extending Proposition 1 to k independent random genomes in \mathbb{G}_n .

To this end, for $A = \{G_1, \dots, G_k\} \subseteq \mathbb{G}_n$, let $\mathcal{B}_A^A = \mathcal{B}_{G_1, \dots, G_k}^A := \mathcal{A}_{G_1, \dots, G_k}$. Then, for any $j = 1, \dots, k$, define

$$\mathcal{B}_{G_1, \dots, G_{j-1}, G_{j+1}, \dots, G_k}^A := \mathcal{A}_{G_1, \dots, G_{j-1}, G_{j+1}, \dots, G_k} \setminus \mathcal{B}_{G_1, \dots, G_k}^A$$

Continuing inductively, for any $I = \{G_{i_1}, \dots, G_{i_r}\} \subset A$, we set

$$\mathcal{B}_I^A = \mathcal{B}_{G_{i_1}, \dots, G_{i_r}}^A := \mathcal{A}_I \setminus \left(\bigcup_{I \subsetneq V} \mathcal{B}_V^A \right).$$

In other words, \mathcal{B}_I^A consists of all adjacencies that are present in every genome in I , but absent from every genome in $A \setminus I$. The sets \mathcal{B}_I^A decompose shared adjacencies according to the subsets of genomes in which they appear.

Let $U_{n,k}, L_{n,k} : (\mathbb{G}_n)^k \rightarrow \mathbb{Z}$ be defined by

$$U_{n,k}(G_1, \dots, G_k) = \sum_{\emptyset \neq I \subseteq [k]} (|I| - 1) |\mathcal{B}_I^A|,$$

and

$$L_{n,k}(G_1, \dots, G_k) = \max_{i \in [k]} \sum_{j \neq i} |\mathcal{A}_{G_i, G_j}|,$$

where in the latter, for each fixed $i \in [k]$, the inner sum runs over all $j \in [k]$ with $j \neq i$, and computes

$$\sum_{j \neq i} |\mathcal{A}_{G_i, G_j}|.$$

The outer maximum then selects the largest value of this sum over all choices of i . Thus, for each genome G_i , we sum the quantities $|\mathcal{A}_{G_i, G_j}|$ over all $j \neq i$, i.e., we add up the sizes of its pairwise adjacency intersections with the other genomes. An adjacency that is shared between G_i and multiple genomes G_j contributes once for each such j , and may therefore be counted multiple times in this sum.

The next result provides explicit upper and lower bounds on how far the median value can fall below the maximal value $(k-1)n$. These bounds depend only on how adjacencies are shared among subsets of the input genomes. The proof is given in Appendix B.

Theorem 1. *For any $k \geq 3$, and any set $A = \{G_1, \dots, G_k\}$ of perfect matchings in \mathbb{G}_n , we have*

$$L_{n,k}(G_1, \dots, G_k) \leq (k-1)n - \mu(A) \leq U_{n,k}(G_1, \dots, G_k). \quad (1)$$

In other words, Theorem 1 shows that the deviation of the median value from its maximum is entirely governed by the shared adjacencies among the input genomes. For random genomes, such shared structures occur only sparsely, leading to tight probabilistic control over the median value.

Now let us apply this result on a set of k independent random perfect matchings $A = \{\mathcal{G}_1, \dots, \mathcal{G}_k\} \subset \mathbb{G}_n$. As mentioned, this is very similar to the case of k non-random perfect matchings with pairwise distance n .

More precisely, for any $I \subseteq [k]$ with $|I| \geq 3$, $\mathcal{A}_I \Rightarrow 0$ as $n \rightarrow \infty$, and $\mathcal{A}_{\mathcal{G}_i, \mathcal{G}_j}$ and $\mathcal{A}_{\mathcal{G}_r, \mathcal{G}_s}$ are asymptotically independent if $\{i, j\} \neq \{r, s\}$. This is immediate when $\{i, j\} \cap \{r, s\} = \emptyset$, since the perfect matchings are independent. To see the case $\{i, j\}, \{i, k\}$, let $\xi_{lr} = \mathbb{1}\{-r, +r\} \in E(\mathcal{G}_l)$. On the other hand, from the invariance of the action of the group of permutations on \mathbb{G}_n , we can assume that $\mathcal{G}_i = o_n$. Then

$$\mathcal{A}_{\mathcal{G}_i, \mathcal{G}_j} = \sum_{r=1}^n \xi_{jr}, \quad \mathcal{A}_{\mathcal{G}_i, \mathcal{G}_k} = \sum_{r=1}^n \xi_{kr}.$$

But since \mathcal{G}_j and \mathcal{G}_k are independent, so is ξ_{jr} and ξ_{ks} for $1 \leq r, s \leq n$.

Then from the Poisson limit for common adjacencies discussed from Appendix A, we deduce, for any $i \in [k]$,

$$\sum_{j \in [k] \setminus \{i\}} |\mathcal{A}_{\mathcal{G}_i, \mathcal{G}_j}| \Rightarrow \text{Poisson}\left(\frac{k-1}{2}\right), \quad (2)$$

and

$$U_{n,k} \Rightarrow \text{Poisson}\left(\frac{k(k-1)}{4}\right). \quad (3)$$

We have readily proved the following result. Let α and β be two independent Poisson random variables, i.e.

$$\alpha \sim \text{Poisson}\left(\frac{k-1}{2}\right), \quad \beta \sim \text{Poisson}\left(\frac{(k-1)(k-2)}{4}\right).$$

Note that $\alpha + \beta$ has the same distributional limit as $U_{n,k}$ in (3). We say a nonnegative integer-valued random variable Z is stochastically dominated by Y , $Z \preceq_{st} Y$, if for every $k \in \mathbb{Z}$, $\mathbb{P}(Z > k) \leq \mathbb{P}(Y > k)$.

The following theorem provides stochastic bounds on the deviation of the median value of k independent random genomes from $(k-1)n$. The proof is given in Appendix B.

Theorem 2. *For a given $k \geq 3$, let $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\} \subseteq \mathbb{G}_n$ be k perfect matchings chosen independently and uniformly at random from \mathbb{G}_n . Then for every $k \in \mathbb{Z}_+$,*

$$\begin{aligned} \mathbb{P}(\alpha > k) &\leq \liminf_{n \rightarrow \infty} \mathbb{P}((k-1)n - \mu(A_n) > k) \\ &\leq \limsup_{n \rightarrow \infty} \mathbb{P}(((k-1)n - \mu(A_n)) > k) \leq \mathbb{P}(\alpha + \beta > k). \end{aligned}$$

Furthermore, $((k-1)n - \mu(A_n))_{n \in \mathbb{Z}_+}$ is tight and hence every sequence of that has a further subsequence that converges in distribution. For every subsequence limit X^* , we have

$$\alpha \preceq_{st} X^* \preceq_{st} \alpha + \beta.$$

The previous theorem shows that the deviation $(k-1)n - \mu(A_n)$ remains stochastically bounded as $n \rightarrow \infty$. In particular, the median value differs from its maximal possible value $(k-1)n$ only by a random quantity of order $O(1)$. The next result formalizes this observation by showing that this deviation becomes negligible under any diverging normalization. The proof is deferred to Appendix B.

Theorem 3. *For any diverging sequence $(a_n)_{n \geq 1}$ of strictly positive numbers, $a_n \rightarrow \infty$, as $n \rightarrow \infty$, we have*

$$\frac{(k-1)n - \mu(A_n)}{a_n} \rightarrow 0, \quad \text{in probability.}$$

While the previous result concerns the median value, it does not yet describe the structure of the median genome itself. The following theorem addresses this question by showing that, with high probability, a median genome introduces only a vanishing proportion of adjacencies not present in the input genomes. The proof is given in Appendix B.

Theorem 4. *Let $k \geq 3$, and let $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$ be k perfect matchings chosen independently and uniformly at random from \mathbb{G}_n . For any $G \in M(A_n)$, then as $n \rightarrow \infty$,*

$$\frac{|\mathcal{A}_G \setminus \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i}|}{n} \xrightarrow{p} 0.$$

Taken together, these results indicate that the bp-median of random genomes is both value-wise and structurally constrained.

3 Position of the median of random genomes

In this section, we study the position of the median genome (perfect matching) of k independent random perfect matchings in \mathbb{G}_n . As shown in the previous section, the median value is asymptotically $(k-1)n$, and the number of adjacencies of the median that do not belong to the union of the input genomes becomes negligible. More precisely, after normalization by any diverging sequence $(a_n)_{n \geq 1}$, this quantity converges to zero in probability.

Before proceeding, we state the guiding question of this section. While the previous section showed that the median value of random genomes is close to its maximum possible value, this does not by itself determine where the median genome is located in genome space. Here we ask whether a median can lie in the interior of the simplex spanned by the input genomes, or whether it must be concentrated near one of the inputs.

Let $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$ be a collection of independent random genomes. Given nonnegative weights x_1, \dots, x_k with $x_1 + \dots + x_k = 1$, our goal is to determine whether there exists a median $\mathcal{G}^{(n)} \in M(A_n)$, whose bp-distance to $\mathcal{G}_i^{(n)}$ is approximately $(1 - x_i)n$, that is $|\mathcal{A}_{\mathcal{G}^{(n)}, \mathcal{G}_i^{(n)}}|/n \rightarrow x_i$, as $n \rightarrow \infty$, in probability. We denote by Δ_k the $(k-1)$ -dimensional probability simplex,

$$\Delta_k := \{(x_1, \dots, x_k) \in \mathbb{R}^k : x_i \geq 0 \text{ for all } i, \text{ and } x_1 + \dots + x_k = 1\}.$$

In fact Δ_k parametrizes all possible normalized overlap profiles between a candidate median and the k input genomes. Points in the interior of Δ_k correspond to medians that take their adjacencies across multiple genomes, while the vertices represent medians that coincide almost entirely with a single input genome.

We will show that this is impossible except in the trivial cases where $x_i = 1$ for exactly one index i and $x_j = 0$ for all $j \neq i$. This means that the approximate medians (and hence exact medians) of k independent random genomes cannot lie far from the “corners”, the input genomes.

We say that a sequence of random variables (Z_n) converges to a random variable Z in L^n , denoted $Z_n \xrightarrow{L^n} Z$, if $\mathbb{E}(|Z_n - Z|^n) \rightarrow 0$, as $n \rightarrow \infty$. It is well known that convergence in L^2 implies convergence in probability, and the latter implies convergence in distribution. In what follows, we exploit this fact through a precise control of certain edge counts in random genomes. A key ingredient in our analysis is a quantitative understanding of how many edges of a random

genome can avoid a prescribed set of gene extremities. The following theorem provides such a result and serves as the main probabilistic input for ruling out interior points of Δ_k . The proof is deferred to Appendix B.

Theorem 5. *Let $m = m_n$ be such that $m/n \rightarrow x$, as $n \rightarrow \infty$. For any given set $R_n \subseteq \pm[n]$ of size $|R_n| = 2m$. Let $\mathcal{G}^{(n)}$ be a perfect matching selected uniformly at random from \mathbb{G}_n , and let $F_n = F_n(R_n)$ be the number of edges in $E(\mathcal{G}^{(n)})$ with both vertices in $\pm[n] \setminus R_n$. Then*

$$\frac{F_n}{n} \xrightarrow{L^2} (1-x)^2, \quad n \rightarrow \infty.$$

Theorem 5 shows that, asymptotically, the number of edges that remain available after fixing a fraction x of gene extremities concentrates sharply around a deterministic value. In particular, once a positive proportion of edges has been fixed, the pool of remaining free edges shrinks quadratically. This concentration phenomenon severely limits the number of edges that can be simultaneously borrowed from multiple random genomes without creating conflicts, and is the key mechanism behind the collapse of feasible overlap profiles to the boundary of the simplex.

Remark 2. As F_n does not depend on the specific choice of R_n but only on its cardinality, we may regard F_n as a function of m , writing $F_n = F_n(m)$.

From the definition of F_n , it is clear that it counts the number of edges in $\mathcal{G}^{(n)}$ whose both ends are free (not used) with respect to R_n . Motivated by this, for a genome $G \in \mathbb{G}_n$ and a set $R \subset \pm[n]$, we say an edge of G is *free* (or with free ends) if both vertices covered by that edge are in $\pm[n] \setminus R$. Theorem 5 counts the number of free edges in a random genome $\mathcal{G}^{(n)}$ with respect to any arbitrary set of gene extremities R_n , with $|R_n| = 2m$.

For a given $x = (x_1, \dots, x_k)$ with $x_1 + \dots + x_k = 1$, $x_i \geq 0$, for $i = 1, \dots, k$ and a set $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$ of perfect matchings chosen independently and uniformly at random from \mathbb{G}_n , we are to see whether it is possible to find a median $\mathcal{G}^{(n)} \in M(A_n)$ so that $|\mathcal{A}_{G, G_i}| \approx \lfloor x_i n \rfloor$.

Let x_1, \dots, x_k be as above, and let $m_1 = m_1(n), \dots, m_k = m_k(n)$ be such that $m_i/n \rightarrow x_i$, as $n \rightarrow \infty$. We are to sample m_1 edges from $\mathcal{G}_1^{(n)}$, m_2 edges from $\mathcal{G}_2^{(n)}$, \dots , and m_k edges from $\mathcal{G}_k^{(n)}$, so that these $m_1 + \dots + m_k$ edges together with the remaining $n - (m_1 + \dots + m_k)$ edges that are not from $\bigcup_{i=1}^k \mathcal{A}_{G_i^{(n)}}$ form a perfect matching (genome) $\mathcal{G} \in \mathbb{G}_n$.

The sampling can be performed on $k!$ different orders. More precisely, any permutation σ on $[k] = \{1, 2, \dots, k\}$ induces an order $\sigma(1), \dots, \sigma(k)$ on $[k]$. Then for any permutation σ on $1, 2, \dots, k$, we sample the adjacencies as follows. We first sample $m_{\sigma(1)}$ edges from $F_n(0) = n$ free edges available in $\mathcal{G}_{\sigma(1)}^{(n)}$, $m_{\sigma(1)} \leq F_n(0) = n$. Denote by $I_{\sigma(1)}$ the set of these sampled edges. Then sample $m_{\sigma(2)}$ edges from $\mathcal{G}_{\sigma(2)}^{(n)}$ that are free with respect to $I_{\sigma(1)}$, therefore we must have $m_{\sigma(2)} \leq F_n(m_{\sigma(1)})$. Continuing in this way, denoting by $I_{\sigma(1)}, \dots, I_{\sigma(r)}$ the edges

sampled from $\mathcal{G}_{\sigma(1)}^{(n)}, \dots, \mathcal{G}_{\sigma(r)}^{(n)}$, we sample $m_{\sigma(r+1)}$ edges from $\mathcal{G}_{\sigma(r+1)}^{(n)}$ which are free with respect to $I_{\sigma(1)} \cup \dots \cup I_{\sigma(r)}$, hence we must have

$$m_{\sigma(r+1)} \leq F_n(\tilde{s}_r^\sigma),$$

where $\tilde{s}_r^\sigma = \tilde{s}_r^\sigma(n) = m_{\sigma(1)} + \dots + m_{\sigma(r)}$ for $r \geq 0$, and $\tilde{s}_0^\sigma = 0$. This yields the constraints, for $r = 0, \dots, k-1$

$$0 \leq m_{\sigma(r+1)} \leq F_n(\tilde{s}_r^\sigma), \quad \tilde{s}_r^\sigma/n \rightarrow s_r^\sigma,$$

as $n \rightarrow \infty$, with $s_0^\sigma = 0$, $s_r^\sigma = x_{\sigma(1)} + \dots + x_{\sigma(r)}$. In particular, $s_k^\sigma = x_{\sigma(1)} + \dots + x_{\sigma(k)} = 1$.

From Theorem 5, we know

$$\frac{F_n(\tilde{s}_r^\sigma)}{n} \xrightarrow{L^2} (1 - s_r^\sigma)^2, \quad n \rightarrow \infty.$$

Hence

$$0 \leq x_{\sigma(r+1)} = \lim_{n \rightarrow \infty} \frac{m_{\sigma(r+1)}}{n} \leq \lim_{n \rightarrow \infty} \mathbb{E} \left(\frac{F_n(\tilde{s}_r^\sigma)}{n} \right) = (1 - s_r^\sigma)^2.$$

Passing to the limit $n \rightarrow \infty$, these sampling constraints translate into deterministic inequalities on the limiting overlap proportions (x_1, \dots, x_k) . Therefore $(x_1, \dots, x_k) \in \Delta_k$ must satisfy

$$x_{\sigma(r+1)} \leq (1 - s_r^\sigma)^2, \quad r = 0, \dots, k-1, \quad (4)$$

for any permutation σ on $[k]$. Inequalities (4) impose strong constraints on the feasible weight vectors (x_1, \dots, x_k) . In particular, they rule out any balanced allocation of overlap across multiple genomes, a fact that will force all feasible limits to lie at the vertices of Δ_k .

Note that By exchangeability of $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$, the order of sampling does not matter. For any given permutation σ on $[k]$, system (4) admits no non-trivial solutions: the only feasible points in Δ_k are the vertices e_1, \dots, e_k . Equivalently, no interior point of the simplex can arise as an asymptotic overlap structure of a median for random genomes. That is, the only solutions of (4) are $x = (x_1, \dots, x_k) = e_i$ for $i = 1, \dots, k$, where e_i denotes the i -th standard basis vector in \mathbb{R}^k .

We say a genome $\mathcal{G}^{(n)}$ is an approximate median for $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$ if, as $n \rightarrow \infty$,

$$\frac{(k-1)n - \sum_{i=1}^k d(\mathcal{G}^{(n)}, \mathcal{G}_i^{(n)})}{n} \rightarrow 0,$$

in probability, or equivalently,

$$\sum_{i=1}^k \frac{|\mathcal{A}_{\mathcal{G}^{(n)}, \mathcal{G}_i^{(n)}}|}{n} \rightarrow 1, \quad \text{in probability.}$$

The arguments above guarantee that with high probability there is no approximate median far from $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$. This conclusion has an important conceptual implication. Although the bp-median is well defined for random genomes, it does not behave like an “average” genome combining features from multiple inputs. Instead, the medians are forced to collapse toward one of the input genomes, both in its total bp-distance and in its adjacency structure. The following theorem is readily obtained from the arguments above, showing that any approximate median must lie asymptotically at a corner of the simplex Δ_k .

Theorem 6. *Let $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$ be k perfect matchings chosen independently and uniformly at random from \mathbb{G}_n , and suppose $\mathcal{G} = \mathcal{G}^{(n)} \in \mathbb{G}_n$ is their approximate median such that as $n \rightarrow \infty$,*

$$\frac{|\mathcal{A}_{\mathcal{G}, \mathcal{G}_i}|}{n} \rightarrow x_i,$$

with $x = (x_1, \dots, x_k) \in \Delta_k$. Then $x \in \{e_1, \dots, e_k\}$.

Remark 3. Theorem 6 implies that approximate medians (and hence exact medians) of a set of random genomes concentrate near the “corners” (i.e. input genomes) with high probability.

4 Anti-medians and a game-theoretic interpretation

In this section we study the midpoints of a set of genomes $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$, sampled independently and uniformly at random from \mathbb{G}_n . The midpoints of $\mathcal{G}_1, \dots, \mathcal{G}_k$ are these perfect matchings (genomes) which are equidistant from the input genomes $\mathcal{G}_1, \dots, \mathcal{G}_k$. In particular we try to find the midpoints with the closest total distance to A_n . These are called anti-medians and are obtained in an attempt to compromise the number of adjacencies shared from the input genomes [8]. We interpret this problem as a game of gene-sharing that is played by k players $\mathcal{G}_1, \dots, \mathcal{G}_k$. This reduces to an optimization problem that we solve and deduce that compromising results in an anti-median behavior. This fully explains the simulation observation regarding the appearance of anti-medians after compromising the gene adjacencies in [8].

Let $A_n = \{\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$ be k perfect matchings picked uniformly and independently at random from \mathbb{G}_n . We frame this as a game in which each genome \mathcal{G}_i plays the role of player i . The players $\mathcal{G}_1, \dots, \mathcal{G}_k$ cooperate. Each player \mathcal{G}_i , for $i = 1, \dots, k$, shares a proportion α_i of its adjacencies, namely I_i , with others. We assume $|I_i| \approx \alpha_i n$. The goal for players is to share their adjacencies such that $\bigcup_{i=1}^k I_i$ be a part of genome G , that is if $\bigcup_{i=1}^k I_i$ constitutes a perfect matching on a subset of $\pm[n]$. Since the perfect matchings are sampled independently and uniformly at random, with high probability their pairwise common adjacencies are few, so we may assume that the I_i are pairwise disjoint, i.e. $I_i \cap I_j = \emptyset$ for $i \neq j$. The payoff for player i is α_i . This is a multiplayer cooperative game with transferable utility, which means side payments among

players are allowed. Therefore the goal for players $\mathcal{G}_1, \dots, \mathcal{G}_k$ is to maximize the utility (payment) function $u(\alpha_1, \dots, \alpha_k) := \alpha_1 + \dots + \alpha_k$, so the problem is to find $(\alpha_1, \dots, \alpha_k)$ such that u is maximized.

It is well known that the optimal solutions of a multiplayer cooperative game with transferable utility lie on the Pareto optimal boundary of the set (domain) of possible payoff vectors (called feasible set).

Definition 1. A feasible payoff vector $\alpha = (\alpha_1, \dots, \alpha_k)$ is Pareto optimal if no feasible vector $\alpha' = (\alpha'_1, \dots, \alpha'_k)$ satisfies $\alpha'_i \geq \alpha_i$ for all i unless $\alpha' = \alpha$.

In other words, no player can gain more without making at least one player worse off. The set of all Pareto-optimal feasible vectors is called the Pareto optimal boundary of the game.

In the sequel, we find the Pareto optimal boundary ϕ_k of our gene-sharing game and show that the minimum payoff on ϕ_k occurs on its intersection with the set $\{(\alpha_1, \dots, \alpha_k) : \alpha_1 = \alpha_2 = \dots = \alpha_k\}$, which is the compromiser of the payoff. We first study the case of two players (two random genomes) and then extend the results to general k .

For two players $\mathcal{G}_1^{(n)}$ and $\mathcal{G}_2^{(n)}$, we may have two sampling mechanisms. Either player 1 first shares $[\alpha_1 n]$ of its adjacencies I_1 , and then player 2 can only share $[\alpha_2 n]$ of its free adjacencies with respect to I_1 , that enforces $\alpha_2 \leq (1 - \alpha_1)^2$; or vice versa, player 2 shares first and player 1 shares next, enforcing $\alpha_1 \leq (1 - \alpha_2)^2$. This means that the set of feasible payoff vectors is

$$P_2 := \{(\alpha_1, \alpha_2) \geq (0, 0) : \alpha_1 \leq (1 - \alpha_2)^2 \text{ or } \alpha_2 \leq (1 - \alpha_1)^2\} \\ = \{(\alpha_1, \alpha_2) \geq (0, 0) : \alpha_1 \leq (1 - \alpha_2)^2\} \cup \{(\alpha_1, \alpha_2) \geq (0, 0) : \alpha_2 \leq (1 - \alpha_1)^2\}.$$

The first condition $\alpha_2 \leq (1 - \alpha_1)^2$ is equivalent to either

$$\alpha_1 \leq \frac{2 - \sqrt{4\alpha_2}}{2} = 1 - \sqrt{\alpha_2} \quad \text{or} \quad \alpha_1 \geq \frac{2 + \sqrt{4\alpha_2}}{2},$$

where the latter is impossible as it contradicts with $\alpha_1 \in [0, 1]$. This implies $\alpha_1 + \sqrt{\alpha_2} \leq 1$. This and the same argument for $\alpha_1 \leq (1 - \alpha_2)^2$ concludes

$$P_2 = \{(\alpha_1, \alpha_2) \geq (0, 0) : \alpha_1 + \sqrt{\alpha_2} \leq 1\} \cup \{(\alpha_1, \alpha_2) \geq (0, 0) : \alpha_2 + \sqrt{\alpha_1} \leq 1\}. \quad (1)$$

This is because the intersection of $\alpha_1 + \sqrt{\alpha_2} = 1$ and $\alpha_2 + \sqrt{\alpha_1} = 1$ occurs at $(\alpha_1, \alpha_2) = (\frac{3-\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2})$ that comes as the intersection of both with the line $\alpha_1 = \alpha_2$, that is the solution to $\alpha_1 + \sqrt{\alpha_2} = 1 = \alpha_2 + \sqrt{\alpha_1}$ (see Figure 1).

We have readily the following result.

Theorem 7. For two-player games with players $\mathcal{G}_1^{(n)}, \mathcal{G}_2^{(n)}$, the Pareto optimal boundary of feasible set of payoff vectors is the function $\phi_2 : [0, 1] \rightarrow [0, 1]$, defined by

$$\alpha_2 = \phi_2(\alpha_1) = \begin{cases} (1 - \alpha_1)^2, & \text{for } \alpha_1 \leq \frac{3 - \sqrt{5}}{2}, \\ 1 - \sqrt{\alpha_1}, & \text{for } \alpha_1 \geq \frac{3 - \sqrt{5}}{2}. \end{cases}$$

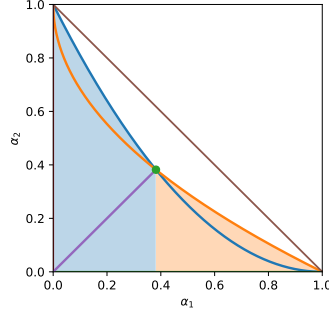


Fig. 1. Feasible payoff region P_2 (shaded) and Pareto boundary for the two-genome game. The boundary consists of the curves $\alpha_2 = (1 - \alpha_1)^2$ and $\alpha_2 = 1 - \sqrt{\alpha_1}$. The diagonal segment $\alpha_1 = \alpha_2$ connects $(0,0)$ to their intersection $(\frac{3-\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2})$, and the line $\alpha_1 + \alpha_2 = 1$ is shown for reference.

The set of feasible vectors P_2 is the area surrounded by (inside and including) $\alpha_1 = 0$, $\alpha_2 = 0$, and ϕ_2 . As mentioned, it is well-known that the optimal solution of gene-sharing game are on ϕ_2 . Otherwise, at least one player can increase its payoff without making the other player worse off. So the first step for players \mathcal{G}_1 and \mathcal{G}_2 is to agree to choose their payoff vector on ϕ_2 . Therefore the optimization problem of finding the best and worst payoff $U(\alpha_1, \alpha_2) = \alpha_1 + \alpha_2$, reduces to:

Optimization Problem 1 (Median). Find $(\alpha_1, \phi_2(\alpha_1))$ with $\alpha_1 \in [0, 1]$ such that the function $\alpha_1 \mapsto \alpha_1 + \phi_2(\alpha_1)$ is maximized.

Optimization Problem 2 (Anti-median). Find $(\alpha_1, \phi_2(\alpha_1))$ with $\alpha_1 \in [0, 1]$ such that the function $\alpha_1 \mapsto \alpha_1 + \phi_2(\alpha_1)$ is minimized.

From the definition, it is clear that the function $\alpha_1 \mapsto \alpha_1 + \phi_2(\alpha_1)$ is strictly decreasing on $[0, \frac{3-\sqrt{5}}{2}]$ and strictly increasing on $[\frac{3-\sqrt{5}}{2}, 1]$. This implies that the solutions to the optimization problems (1) are $(\alpha_1, \alpha_2) = (0, 1)$ and $(\alpha_1, \alpha_2) = (1, 0)$. In other words the function $\alpha_1 \mapsto \alpha_1 + \phi_2(\alpha_1)$, attains its maximum at $\alpha_1 = 0$ (i.e. $(\alpha_1, \alpha_2) = (0, 1)$) and $\alpha_1 = 1$ (i.e. $(\alpha_1, \alpha_2) = (1, 0)$) with maximum value $U(0, 1) = U(1, 0) = 1$ that corresponds to the approximate medians tending to the corners (input genomes $\mathcal{G}_1^{(n)}$ and $\mathcal{G}_2^{(n)}$).

On the other hand, the unique solution to the optimization problem (2) is $(\alpha_1, \alpha_2) = (\frac{3-\sqrt{5}}{2}, \frac{3-\sqrt{5}}{2})$. In other words, the function $\alpha_1 \mapsto \alpha_1 + \phi_2(\alpha_1)$ attains its minimum on $\alpha_1 = \frac{3-\sqrt{5}}{2}$ (i.e. $(\alpha_1, \alpha_2) = ((3 - \sqrt{5})/2, (3 - \sqrt{5})/2)$) which corresponds to full compromise and anti-median.

Furthermore, as $\alpha_1 \mapsto \alpha_1 + \phi_2(\alpha_1)$ is decreasing on $\alpha_1 \leq (3 - \sqrt{5})/2$ and increasing on $\alpha_1 \geq (3 - \sqrt{5})/2$, we can conclude that the more \mathcal{G}_1 and \mathcal{G}_2 compromise, the less the genome G constructed from their shared adjacencies behaves like a median, i.e. the more G behaves like an anti-median. This fully explains the observations from the simulation studies in [9, 8] for $k = 2$.

The same idea extends to k players, although the Pareto optimal boundary will become a hyper-surface ϕ_k on \mathbb{R}_+^k .

Recall that the total payoff is $u(\alpha_1, \dots, \alpha_k) = \alpha_1 + \dots + \alpha_k$. Denote by P_k the set of feasible payoff vectors for $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$. Then $P_k = \cup_{\sigma} P_k^{\sigma}$, where the union is over all permutations σ of $[k]$, and P_k^{σ} is the set of all $\alpha = (\alpha_1, \dots, \alpha_k) \in \mathbb{R}_+^k$, with $\alpha_1, \dots, \alpha_k \leq 1$, for which

$$\alpha_{\sigma(r+1)} \leq \left(1 - \sum_{t=1}^r \alpha_{\sigma(t)}\right)^2 \quad r = 1, \dots, k-1.$$

As the optimal payoff vectors are located on ϕ_k , first $\mathcal{G}_1, \dots, \mathcal{G}_k$ agree to take their payoff vector on ϕ_k , and then they try to optimize it. Having themselves restricted to ϕ_k , the maximum and minimum payoff on ϕ_k can be stated and solved similarly.

It concludes that the maximum value of u on ϕ_k occurs at payoff vectors $\alpha = (\alpha_1, \dots, \alpha_k) = e_i$, for $i = 1, \dots, k$, where as before $e_i \in \mathbb{R}^k$ denotes the i -th standard basis vector. This corresponds to the approximate median of $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$ that all tends to the corners.

On the other hand the minimum value of u on ϕ_k occurs at the unique point on the intersection of ϕ_k and the diagonal of \mathbb{R}_+^k , i.e.

$$\{v^*\} = \phi_k \cap \{(\alpha_1, \dots, \alpha_k) \in \mathbb{R}_+^k : \alpha_1 = \alpha_2 = \dots = \alpha_k\}.$$

Letting $v^* = (\theta(k), \dots, \theta(k))$, equivalently $\alpha_1 = \alpha_2 = \dots = \alpha_k = \theta(k)$, we have

$$\theta(k) \leq (1 - r\theta(k))^2,$$

for $r = 0, \dots, k-1$. Since the right-hand side decreases as r increases, the strongest condition is when $r = k-1$. Therefore, to have v^* on the Pareto optimal boundary, we must have equality, i.e.

$$\theta(k) = (1 - (k-1)\theta(k))^2.$$

Note that the larger root of the last equation is not admissible in the present setting, as it does not lie in the set of feasible payoff vectors. Therefore, $\theta(k)$ is the smaller root, which is given by

$$\theta(k) = \frac{2k-1 - \sqrt{4k-3}}{2(k-1)^2}.$$

The total payoff is

$$u(v^*) = k\theta(k) = \frac{k(2k-1 - \sqrt{4k-3})}{2(k-1)^2},$$

and the total distance of v^* to $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$ is approximately $kn(1 - \theta(k))$.

This solution corresponds to the anti-median of $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$. Between these two scenarios, one can see that the total payoff decreases as we move on lines between e_i to v^* , for $i = 1, \dots, k$. This fully explains the appearance of anti-medians after compromising the number of gene adjacencies taken from $\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}$ and shows why the total distance increases as we increase compromization.

5 Discussion

We have shown that for independently and uniformly sampled signed multichromosomal genomes, breakpoint medians asymptotically lie at the corners: with high probability, any median genome shares almost all of its adjacencies with a single input genome. This rigorously establishes the Haghghi–Sankoff conjecture in this setting and provides a unified explanation for earlier heuristic and simulation-based observations.

This phenomenon is driven by the strong constraints inherent in the signed multichromosomal model. Although breakpoint medians are well defined, the set of feasible medians is highly restricted: balanced overlap profiles are ruled out, and any approximate median is forced to draw almost all of its adjacencies from a single input genome. As a consequence, strategies that attempt to combine adjacencies evenly from multiple genomes necessarily increase the total distance.

The game-theoretic framework developed here makes this mechanism explicit, explaining why increasing compromise along the Pareto boundary leads to increased total bp-distance to the input genomes and to the emergence of anti-medians.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bryant, D.: The complexity of the breakpoint median problem. Centre de recherches mathématiques (1998)
2. Caprara, A.: The reversal median problem. *INFORMS Journal on Computing* **15**(1), 93–113 (2003)
3. Chen, L.H.Y.: Poisson approximation for dependent trials. *The Annals of Probability* **3**(3), 534–545 (1975). <https://doi.org/10.1214/aop/1176996303>
4. Feijão, P., Meidanis, J.: SCJ: a variant of breakpoint distance for which sorting, genome median and genome halving problems are easy. In: *International Workshop on Algorithms in Bioinformatics*. pp. 85–96. Springer (2009)
5. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: *Combinatorics of genome rearrangements*. The MIT Press (2009)
6. Haghghi, M., Sankoff, D.: Medians seek the corners, and other conjectures. *BMC Bioinformatics* **13**(19), S5 (2012)
7. Jamshidpey, A., Jamshidpey, A., Sankoff, D.: Sets of medians in the non-geodesic pseudometric space of unsigned genomes with breakpoints. *BMC Genomics* **15**(6), S3 (2014)
8. Larlee, C.A., Brandts, A., Sankoff, D.: Compromise or optimize? the breakpoint anti-median. *BMC bioinformatics* **17**(Suppl 18), 473 (2016)
9. Larlee, C.A., Zheng, C., Sankoff, D.: Near-medians that avoid the corners; a combinatorial probability approach. *BMC Genomics* **15**(6), S1 (2014)
10. Sankoff, D., Blanchette, M.: The median problem for breakpoints in comparative genomics. *Computing and Combinatorics* pp. 251–263 (1997)

11. Sankoff, D., Sundaram, G., Kececioğlu, J.: Steiner points in the space of genome rearrangements. *International Journal of Foundations of Computer Science* **7**(01), 1–9 (1996)
12. da Silva, P.H., Jamshidpey, A., Sankoff, D.: Sampling gene adjacencies and geodesic points of random genomes. In: *RECOMB International Workshop on Comparative Genomics*. pp. 189–210. Springer (2024)
13. da Silva, P.H., Jamshidpey, A., Sankoff, D.: Identifying breakpoint median genomes: A branching algorithm approach. In: *25th International Conference on Algorithms for Bioinformatics (WABI 2025)*. pp. 18–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik (2025)
14. Stein, C.: A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume II: Probability Theory*. pp. 583–602. University of California Press (1972)
15. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics* **10**(1), 120 (2009)
16. Xu, A.W.: The median problems on linear multichromosomal genomes: Graph representation and fast exact solutions. *Journal of Computational Biology* **17**(9), 1195–1211 (2010)
17. Xu, W., Alain, B., Sankoff, D.: Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases. *Bioinformatics* **24**(16), i146–i152 (2008)
18. Zanetti, J.P.P., Biller, P., Meidanis, J.: Median approximations for genomes modeled as matrices. *Bulletin of Mathematical Biology* **78**, 786–814 (2016)

A Bp-distance of two random genomes

In this part, we will obtain a quantitative control on the approximation error for the distance of two independent random genomes (perfect matchings). Indeed, it is not hard to show a stronger convergence. To see this, recall that the total variation distance for two random variables Y and Z with a common state space S is defined by

$$\|\mathcal{L}(Y) - \mathcal{L}(Z)\|_{TV} = \sup_{A \subseteq S} |\mathbb{P}(Y \in A) - \mathbb{P}(Z \in A)|,$$

where $\mathcal{L}(Y)$ and $\mathcal{L}(Z)$ denote the distributions of random variables Y and Z , respectively. When $S \subseteq \mathbb{Z}_+$, the last formula reduces to

$$\|\mathcal{L}(Y) - \mathcal{L}(Z)\|_{TV} = \frac{1}{2} \sum_{i \in S} |\mathbb{P}(Y = i) - \mathbb{P}(Z = i)|.$$

We show that $|\mathcal{A}_{\mathcal{G}_n, \mathcal{G}'_n}|$ converges to Poisson(1/2) in total variation topology, i.e., as $n \rightarrow \infty$,

$$\|\mathcal{L}(|\mathcal{A}_{\mathcal{G}_n, \mathcal{G}'_n}|) - \mathcal{L}(\text{Poisson}(1/2))\|_{TV} \rightarrow 0.$$

As we discussed earlier, this is equivalent to showing that, as $n \rightarrow \infty$,

$$\|\mathcal{L}(|\mathcal{A}_{\mathcal{G}_n, o_n}|) - \mathcal{L}(\text{Poisson}(1/2))\|_{TV} \rightarrow 0. \quad (5)$$

For simplicity, let $C_n := |\mathcal{A}_{\mathcal{G}_n, o_n}|$. We proceed as follows. Let \mathcal{G}_n be a random perfect matching in \mathbb{G}_n , and let

$$\xi_{in} = \mathbb{1}_{\{-i, +i\} \in E(\mathcal{G}_n)},$$

where $\mathbb{1}$ is the indicator function, i.e. $\xi_{in} = 1$ if $\{-i, +i\} \in E(\mathcal{G}_n)$ and $\xi_{in} = 0$ otherwise. Each indicator ξ_{in} records whether the i -th adjacency of the reference genome is preserved in the random genome. We have

$$\mathbb{E}(\xi_{in}) = \frac{1}{2n-1}, \quad \mathbb{E}(\xi_{in}\xi_{jn}) = \frac{1}{(2n-1)(2n-3)}, \quad i \neq j.$$

Therefore,

$$\text{Var}(\xi_{in}) = \frac{1}{2n-1} \left(\frac{2n-2}{2n-1} \right), \quad \text{Cov}(\xi_{in}, \xi_{jn}) = \frac{2}{(2n-1)^2(2n-3)}.$$

On the other hand,

$$C_n = \sum_{i=1}^n \xi_{in}.$$

Hence, as $n \rightarrow \infty$,

$$\mathbb{E}(C_n) = \frac{n}{2n-1} \rightarrow \frac{1}{2},$$

and

$$\text{Var}(C_n) = \frac{n(2n-2)}{(2n-1)^2} + 2(n^2-n) \left(\frac{2}{(2n-1)^2(2n-3)} \right) \rightarrow \frac{1}{2}.$$

In addition, for any $n \in \mathbb{N}$, the random variables $\xi_{1n}, \dots, \xi_{nn}$ are exchangeable. That is, for any permutation $\sigma \in S_n$, where S_n is the group of permutations on $\{1, \dots, n\}$, and any $(j_{1n}, j_{2n}, \dots, j_{nn}) \in \{0, 1\}^n$, we have

$$\mathbb{P}(\xi_{1n} = j_{1n}, \dots, \xi_{nn} = j_{nn}) = \mathbb{P}(\xi_{\sigma(1)n} = j_{1n}, \dots, \xi_{\sigma(n)n} = j_{nn}).$$

Thus, by the Stein–Chen method [14, 3], we have

$$\|\mathcal{L}(C_n) - \mathcal{L}(\text{Poisson}(n/(2n-1)))\|_{TV} \leq \frac{1 - e^{-n/(2n-1)}}{n/(2n-1)} \left(\text{Var}(C_n) - \frac{n}{2n-1} + \frac{2n}{2n-3} \right) \xrightarrow{n \rightarrow \infty} 0. \quad (6)$$

Hence (5) follows from (2) and triangle inequality for the total variation distance. It is well-known that convergence in total variation implies convergence in distribution

$$C_n \Rightarrow \text{Poisson}(1/2), \quad n \rightarrow \infty. \quad (7)$$

That is,

$$\mathbb{P}(C_n = i) \rightarrow \frac{(1/2)^i e^{-1/2}}{i!}, \quad i \in \mathbb{Z}_+.$$

If we normalize C_n by a sequence of positive real numbers $(a_n)_{n \geq 1}$ such that $a_n \rightarrow 0$, then from (5) or (7) we obtain

$$\frac{C_n}{a_n} \rightarrow 0, \quad \text{in probability.} \quad (8)$$

Recall that, for a sequence of random variables Z_n , $n \in \mathbb{N}$, we say Z_n converges in probability to a random variable Z , denoted by $Z_n \xrightarrow{P} Z$, if for any $\varepsilon > 0$, as $n \rightarrow \infty$,

$$\mathbb{P}(|Z_n - Z| > \varepsilon) \rightarrow 0.$$

B Proofs

B.1 Proof of Proposition 1

Proof. We follow the proof from [7], for unsigned genomes. For $G \in M(A)$, suppose $\mathcal{A}_G \setminus \bigcup_{i=1}^k \mathcal{A}_{G_i} \neq \emptyset$. Let $e \in \mathcal{A}_G \setminus \bigcup_{i=1}^k \mathcal{A}_{G_i}$. Since $e \notin \mathcal{A}_{G_i}$ for every i , this edge contributes to none of the intersections \mathcal{A}_{G, G_i} , hence

$$\sum_{i=1}^k |\mathcal{A}_{G, G_i}| \leq n - 1.$$

Therefore

$$\sum_{i=1}^k d(G, G_i) = \sum_{i=1}^k (n - |\mathcal{A}_{G, G_i}|) = kn - \sum_{i=1}^k |\mathcal{A}_{G, G_i}| > kn - n = (k - 1)n.$$

On the other hand, for each $i = 1, 2, \dots, k$ we have

$$\sum_{j \neq i} d(G_i, G_j) = (k - 1)n,$$

so G cannot minimize the total distance, a contradiction with the fact that $G \in M(A)$. This implies $\mathcal{A}_G \setminus \bigcup_{i=1}^k \mathcal{A}_{G_i} = \emptyset$. Therefore $\sum_i |\mathcal{A}_{G, G_i}| = n$, and from the above,

$$\sum_{i=1}^k d(G, G_i) = (k - 1)n = \mu(A). \quad \blacksquare$$

B.2 Proof of Theorem 1

Proof. For a given set A of genomes, the best scenario for a median, i.e. the smallest possible median value is attained, happens if not only is every adjacency of every median $G \in M(A)$ an adjacency of a genome in A , but also

$$\bigcup_{j \neq i} \mathcal{A}_{G_i, G_j} \subseteq \mathcal{A}_G.$$

From definition, then \mathcal{A}_G can be partitioned into $\{\mathcal{B}_{G_i G_j} : 1 \leq i < j \leq k\} \cup \{\mathcal{B}_{G G_i} : i \in [k]\}$. Hence

$$\begin{aligned} \sum_{i=1}^k d(G, G_i) &= \sum_{i=1}^k \left(n - |\mathcal{B}_{G G_i}| - \sum_{\substack{i \in I \\ |I| \geq 2}} |I| |\mathcal{B}_I| \right) \\ &= kn - \left(\sum_{i=1}^k |\mathcal{B}_{G G_i}| + \sum_{\substack{i \in I \\ |I| \geq 2}} |\mathcal{B}_I| \right) - \sum_{\substack{i \in I \\ |I| \geq 2}} (|I| - 1) |\mathcal{B}_I| \\ &= (k-1)n - U_{n,k}(G_1, \dots, G_k). \end{aligned}$$

Thus the right hand side of (1). For the left side of (1), note that for any $G \in M(A)$,

$$\mu(A) = \sum_{j=1}^k d(G, G_j) \leq \sum_{j \neq i} d(G_i, G_j) \leq \min_i \sum_{j \neq i} (n - |\mathcal{A}_{G_i, G_j}|),$$

therefore

$$(k-1)n - \mu(A) \geq L_{n,k}(G_1, \dots, G_k). \quad \blacksquare$$

B.3 Proof of Theorem 2

Proof. The first part follows (2), (3) and applying Theorem 1. Note that $((k-1)n - \mu(A_n))_{n \in \mathbb{Z}_+}$ is bounded between two sequences converging to α and $\alpha + \beta$, respectively. Therefore we can see that $((k-1)n - \mu(A_n))_{n \in \mathbb{Z}_+}$ is tight. Therefore, from Helly's selection theorem (Billingsley ...), every subsequence of that has a further subsequence that converges in distribution. Hence, for any subsequence limit X_n , namely X^* , we will have

$$\mathbb{P}(\alpha > k) \leq \mathbb{P}(X^* > k) \leq \mathbb{P}(\alpha + \beta > k). \quad \blacksquare$$

B.4 Proof of Theorem 3

Proof. Let $X_n := (k-1)n - \mu(A_n) \geq 0$. By the previous theorem, the sequence $(X_n)_{n \geq 1}$ is tight. Hence, for every $\varepsilon > 0$, there exists $M < \infty$ such that

$$\sup_{n \geq 1} \mathbb{P}(X_n > M) < \varepsilon.$$

Let $(a_n)_{n \geq 1}$ be any diverging sequence of positive numbers with $a_n \rightarrow \infty$. Fix $\delta > 0$. Choose N such that for all $n \geq N$ we have $\delta a_n \geq M$. Then, for $n \geq N$,

$$\mathbb{P}\left(\frac{X_n}{a_n} > \delta\right) = \mathbb{P}(X_n > \delta a_n) \leq \mathbb{P}(X_n > M) \leq \sup_{m \geq 1} \mathbb{P}(X_m > M) < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, it follows that

$$\frac{X_n}{a_n} \xrightarrow{p} 0, \quad n \rightarrow \infty. \quad \blacksquare$$

B.5 Proof of Theorem 4

Proof. Fix n and write $A_n = \{\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}\}$. Let $G \in M(A_n)$ and set

$$t_n := |\mathcal{A}_G \setminus \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}|.$$

For each adjacency (edge) $e \in \mathcal{A}_G$, define

$$c(e) := |\{i \in [k] : e \in \mathcal{A}_{\mathcal{G}_i^{(n)}}\}| \in \{0, 1, \dots, k\}.$$

Then $c(e) = 0$ if and only if $e \notin \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}$. Moreover,

$$\sum_{i=1}^k |\mathcal{A}_{G, \mathcal{G}_i^{(n)}}| = \sum_{e \in \mathcal{A}_G} c(e) = \sum_{e \in \mathcal{A}_G \cap \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}} c(e).$$

Since $c(e) \geq 1$ on the intersection, we may write

$$\sum_{e \in \mathcal{A}_G \cap \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}} c(e) = (n - t_n) + \sum_{e \in \mathcal{A}_G \cap \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}} (c(e) - 1).$$

On the other hand,

$$(k-1)n - \mu(A_n) = \sum_{i=1}^k |\mathcal{A}_{G, \mathcal{G}_i^{(n)}}| - n,$$

since $\mu(A_n) = \sum_{i=1}^k d(G, \mathcal{G}_i^{(n)}) = kn - \sum_{i=1}^k |\mathcal{A}_{G, \mathcal{G}_i^{(n)}}|$. Combining the last two displays gives

$$(k-1)n - \mu(A_n) = -t_n + \sum_{e \in \mathcal{A}_G \cap \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}} (c(e) - 1),$$

and hence

$$t_n = \sum_{e \in \mathcal{A}_G \cap \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}} (c(e) - 1) - ((k-1)n - \mu(A_n)) \leq \sum_{e \in \bigcup_{i=1}^k \mathcal{A}_{\mathcal{G}_i^{(n)}}} (c(e) - 1).$$

The right-hand side equals $U_{n,k}(\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)})$ by the definition of $U_{n,k}$ (equivalently, by the decomposition into the \mathcal{B}_I^A sets). Therefore,

$$0 \leq t_n \leq U_{n,k}(\mathcal{G}_1^{(n)}, \dots, \mathcal{G}_k^{(n)}).$$

Finally, by (3) we have $U_{n,k} \Rightarrow \text{Poisson}\left(\frac{k(k-1)}{4}\right)$, hence $(U_{n,k})_{n \geq 1}$ is tight. For any $\varepsilon > 0$ choose M such that $\sup_{n \geq 1} \mathbb{P}(U_{n,k} > M) < \varepsilon$. Then, for any $\delta > 0$ and all n large enough that $\delta n \geq M$,

$$\mathbb{P}\left(\frac{t_n}{n} > \delta\right) \leq \mathbb{P}\left(\frac{U_{n,k}}{n} > \delta\right) = \mathbb{P}(U_{n,k} > \delta n) \leq \mathbb{P}(U_{n,k} > M) < \varepsilon.$$

Since ε is arbitrary, this implies $t_n/n \xrightarrow{P} 0$. ■

B.6 Proof of Theorem 5

Proof. Consider the n edges of $\mathcal{G}^{(n)}$ and order them in random $\tilde{e}_1, \dots, \tilde{e}_n$. For an edge e , let $\text{end}(e)$ be the two vertices incident to it. For $i = 1, \dots, n$, let

$$\xi_i^{(n)} = \mathbb{1}_{\{\text{end}(\tilde{e}_i) \cap R_n = \emptyset\}}.$$

Then $F_n = \sum_{i=1}^n \xi_i^{(n)}$. We can compute the expected value and variance of F_n . Write

$$\mathbb{E}(F_n) = \sum_{i=1}^n \mathbb{P}(\text{end}(\tilde{e}_i) \cap R_n = \emptyset) = \sum_{i=1}^n \frac{(2n-2m)(2n-2m-1)}{2n(2n-1)}.$$

So

$$\mathbb{E}\left(\frac{F_n}{n}\right) \rightarrow (1-x)^2.$$

Also,

$$\begin{aligned} \sum_{i=1}^n \text{Var}(\xi_i^{(n)}/n) &= \frac{n}{n^2} (\mathbb{E}(\xi_i) - (\mathbb{E}(\xi_i))^2) = \frac{1}{n} \mathbb{E}(\xi_i)(1 - \mathbb{E}(\xi_i)) \\ &= \frac{(2n-2m)(2n-2m-1)}{n(2n(2n-1))} \left(1 - \frac{(2n-2m)(2n-2m-1)}{2n(2n-1)}\right) \rightarrow 0, \end{aligned}$$

and

$$\begin{aligned} 2 \sum_{i < j} \text{Cov}(\xi_i^{(n)}/n, \xi_j^{(n)}/n) &= \frac{(n^2 - n)}{n^2} \times \\ &\left(\frac{(2n-2m)(2n-2m-1)(2n-2m-2)(2n-2m-3)}{(2n)(2n-1)(2n-2)(2n-3)} - \frac{(2n-2m)^2(2n-2m-1)^2}{2n(2n-1)} \right) \\ &= \frac{(n-1)(2n-2m)(2n-2m-1)}{n(2n)(2n-1)} \times \\ &\left(\frac{(2n-2m-2)(2n-2m-3)}{(2n-2)(2n-3)} - \frac{(2n-2m)(2n-2m-1)}{2n(2n-1)} \right) \end{aligned}$$

where the right-hand side converges to 0, as $n \rightarrow \infty$. Thus, $\lim_{n \rightarrow \infty} \text{Var}(F_n) = 0$.

Since $\mathbb{E}(F_n) \rightarrow (1-x)^2$ and $\text{Var}(F_n) \rightarrow 0$, as $n \rightarrow \infty$, from Chebyshev's inequality we conclude that F_n converges to $(1-x)^2$ in L^2 . ■

Remark 4. One can obtain the asymptotics for the normalized number of edges in $\mathcal{G}^{(n)}$ with both ends in R_n , denoted by $\tilde{F}_n = \tilde{F}_n(R_n)$, or the normalized number of edges in $\mathcal{G}^{(n)}$ with exactly one end in R_n and the other end in $\pm[n] \setminus R_n$, denoted by $\tilde{F}'_n = \tilde{F}'_n(R_n)$. For \tilde{F}_n , we apply Theorem 5 to the set $\pm[n] \setminus R_n$ instead of R_n , and obtain

$$\frac{\tilde{F}_n}{n} \xrightarrow{L^2} x^2, \quad n \rightarrow \infty.$$

For \widetilde{F}'_n , we note that $\widetilde{F}'_n = n - \widetilde{F}_n - F_n$, hence

$$\frac{\widetilde{F}'_n}{n} \xrightarrow{L^2} 1 - x^2 - (1 - x)^2 = 2x(1 - x).$$