Large harmonious sets of non-crossing edges for n randomly labeled vertices in convex position

József Balogh*
Boris Pittel†
Gelasio Salazar ‡

April 26, 2005

Abstract

Consider a set S of points in the plane in general position, where each point has an integer label from $\{0, 1, \ldots, n-1\}$. This naturally induces a labeling of the edges: each edge (i, j) is assigned the label $i + j$, modulo n. In the spirit of harmonious graphs, we propose algorithms for finding (hopefully) large non–crossing *harmonic* matchings or paths, i. e. the matchings or paths in which no two edges have the same label. When the point labels are chosen uniformly at random, and independently of each other, our matching algorithm with high probability (w.h.p.) delivers a nearly–perfect matching, a matching of size $n/2 - O(n^{1/3} \ln n)$. We show that, in sharp contrast, a near-perfect path is unlikely: w.h.p. the length of the longest path is below $0.96n$. Our empirically best path algorithm seems to consistently build a path of length above $0.78n$, and the likely path length for our second best algorithm is provably above $0.66n$.

Keywords: Graceful labeling, harmonious labeling, noncrossing, harmonic graph, convex position, point sets, matching, path

AMS Subject Classification: 05C70, 52B55, 05C85, 68W40

1 Introduction

We are motivated by the concepts of graceful labelings and harmonious graphs introduced by Graham and Sloane [5] (see [4] for a comprehensive survey on these problems). Our interest

*Dept. of Mathematics, The Ohio State University, Columbus, OH 43210, USA. Supported by NSF Grant DMS–0302804.

†Dept. of Mathematics, The Ohio State University, Columbus, OH 43210, USA. Supported by NSF Grant DMS–0406024.

‡Instituto de Fisica, Universidad Autonoma de San Luis Potosi, San Luis Potosi, SLP, Mexico 78000. Supported by FAI–UASLP. Work was partly done while at OSU.

is in the problem of existence of large substructures (subsets of edges or subgraphs) such that all the edges involved have different labels. Typically, an edge label is a function of the labels of the endvertices, e.g. the absolute value of their difference (graceful labelings), or their sum modulo some n (harmonious graphs). There is another natural rule for assigning the edge labels: an edge gets a label equal to the *product* of its endpoints labels, modulo n. Curiously, for n prime we have basically the same model as the multiplicative group on $\{1, \dots, n-1\}$ is isomorphic to the additive group on $\{0, 1, \dots, n-2\}$. For n composite, some probabilistic-number theoretic issues are likely to arise.

For the point set in the plane it is natural to seek large substructures (paths, matchings) that meet certain geometric conditions. One popular *non–crossing* condition requires that no two edges in the substructure cross each other. For a sample of diverse results in this area of combinatorial geometry we refer the reader to [1, 2, 6, 7, 8, 9, 16].

To describe the results of this paper, we need some terminology and notations. Following [3], let S be a set of points in the plane in a convex position. Assume that each point has an integer label from $\{0, \dots, n-1\}$. If p, q are distinct points (also called *vertices*) in S, then we let (p, q) denote the straight segment (or *edge*) that has p and q as its endvertices. This naturally induces a (complete) *geometric graph* G_S. In general, we let $E(K)$ denote the set of edges of a graph K. A subset E' of $E(G_S)$ is *non–crossing* if no two edges in E' intersect in a point other that a common endvertex. A subgraph H of G_S is *non–crossing* if $E(H)$ is non–crossing.

As for the edge labels, we use the sum rule; it assigns to each edge (p, q) a number equal to the sum of labels of p and q modulo n. One such rule assigns to each edge the sum (modulo n) of the labels of its endpoints. In this geometric setting, the central problem is to find conditions for existence of large non–crossing subgraphs whose edge labels are all distinct.

While [3] dealt exclusively with the worst–case instances of the labeled set S, our goal is to study the average (likely) case behavior under assumption that the labels of points in S are random. More specifically, we assume that each of the n points is labeled with an integer drawn uniformly at random from $\{0, 1, 2, \dots, n-1\}$, independently of all other labels. We pose the following questions.

Question 1 *How many edges are there typically in a maximum size harmonic non–crossing matching in G_S?*

Question 2 *How many edges are there typically in a maximum size harmonic non–crossing path in G_S?*

In our opinion, we have found a satisfying answer to Question 1. We propose a greedy matching algorithm (HMatching) that w.h.p. delivers a matching of size $n/2 - O(n^{1/3} \ln n)$—a nearly perfect matching, as the number of unmatched vertices is w.h.p. merely of order $n^{1/3} \ln n$. Thus the maximum matching number is $n/2 - O(n^{1/3} \ln n)$ at least. (For the Erdős-Rényi random graph with $n/2$ edges, i. e. in the critical stage, the core vertices of degree more than, or equal to 3 are typically incident to $O(n^{1/3})$ edges, Łuczak et al [10]. Similarity between the numbers in both schemes is hardly more than coincidental though.) For an arbitrary starting point the probability that the resulting matching is perfect is not too small, of order $\Omega(n^{-1/3}\ln^{-1} n)$, i. e. the expected number of “lucky” starting points is $\Omega(n^{2/3}\ln^{-1} n)$. We conjecture that the number itself is likely to be that large as well, so that w.h.p. there exists a perfect matching! In Section 2 we present HMatching, and in Section 3 we give the experimental results that allowed us to predict the likely behavior of the algorithm. In Section 4 we provide a rigorous analysis which confirms—within the logarithmic factors— the conjectured bounds. In Section 5 we briefly discuss the related problem in which the point labels form a random permutation of $(0, 1, 2, \dots, n-1)$, rather than being strictly independent of each other.

Somewhat unexpectedly Question 2 is inherently harder to answer fully. In Section 6 we describe two greedy path algorithms, one rather naive, another quite elaborate. The first algorithm w.h.p. delivers a path of length asymptotic, in probability, to $(1-e^{-1})n \approx 0.63n$. The computer runs indicate that the second algorithm is considerably more efficient, consistently delivering a path of length close to $0.76n$. The experiments and a caricature model of this algorithm compel us to conjecture that the algorithm finds a path of length asymptotic, in probability, to $n(1-e^{-2})/(1+e^{-2}) \approx 0.761n$. In Section 6.4 we describe HPath, a compromise algorithm, whose typical performance puts it between the first two algorithms. In Section 7 we show that w.h.p. this algorithm finds a path of length $0.66n$ at least. These results signal that, in sharp contrast with the matching problem, the longest path is not likely to contain $n-o(n)$ vertices. And indeed, using a counting (nonalgorithmic) argument we show (Section 8) that w.h.p. L_n, the length of the longest path, falls below $0.96n$. We conjecture that L_n/n converges, in probability, to a constant between 0.79 and 0.96.

We conclude with the following question.

Quastion 3 *How many edges are there typically in a maximum size harmonic non–crossing tree or forest in G_S?*

Going out on a limb, we conjecture that w.h.p. the maximum tree size is $n-1$, so that the maximum tree spans all n vertices.

2 HMatching: the algorithm

Naturally, the first step in our quest for a satisfactory answer to Question 1 was to come up with an algorithm that would yield, in computer simulations, large harmonic non–crossing matchings.

After several attempts, we settled on a reasonably simple algorithm that consistently ended up with very large matchings in the computer experiments. We call this algorithm HMatching.

Our basic assumption is that the n points that comprise the set S are in convex position, so that all the points are on the boundary of the convex hull of S. No relevant geometrical information is lost if we assume that all the points lie on a circle. Therefore, we may denote the points as $p_0, p_1, \dots, p_{n-1}$, according to the cyclic (counter-clockwise) order in which they appear on the circle. Further each point p_i gets a label $A[i]$, and the n labels are drawn independently from the uniform distribution on $\{0, 1, \dots, n-1\}$. Given the point labels, each edge (p_i, p_j) gets the label $A[i, j] :\equiv A[i] + A[j] \pmod{n}$.

HMatching takes as input an array $(A[0], A[1], A[2], \dots, A[n-1])$ and its output is a (non–crossing, harmonic) matching on S. At each step we have a current matching, both non–crossing and harmonic, to which we add a new edge to get a larger matching that meets the same requirements. Formally, we maintain the current matching $\mathcal{M}$ as a collection of ordered pairs (i, j) with $i < j$, where (i, j) represents (p_i, p_j). Clearly the edge set $\mathcal{M}$ satisfies the following conditions:

(a) if (i, j) and (i', j') are different pairs in $\mathcal{M}$, then $\{i, j\} \cap \{i', j'\} = \emptyset$ ($\mathcal{M}$ is a matching);

(b) if (i, j) and (i', j') are different pairs in $\mathcal{M}$, then $A[i] + A[j] \not\equiv A[i'] + A[j'] \pmod{n}$ ($\mathcal{M}$ is harmonic);

(c) if (i, j) and (i', j') are different pairs in $\mathcal{M}$, with $i < i'$, then either $i < j < i' < j'$ or $i < i' < j' < j$ ($\mathcal{M}$ is non–crossing).

The pseudocode for HMatching is the following.:

Input : An array $(A[0], A[1], \dots, A[n-1])$, such that $A[i] \in \{0, 1, \dots, n-1\}$ for every i.

Output : The size of a set $\mathcal{M}$ of pairs (i, j), with $i < j$, that satisfies (a), (b), and (c) above.

Procedure :

```
$S = \emptyset$;  $\mathcal{M} = \emptyset$;  $L = \emptyset$;  $k = 0$
while $k \leq n - 1$
    do
        if $S \neq \emptyset$
            then if $A[\max S] + A[k] \pmod{n} \notin L$
                then $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\max S, k)\}$
                     $L \leftarrow L \cup \{A[\max S] + A[k] \pmod{n}\}$
                     $S \leftarrow S \setminus \{\max S\}$
                else $S \leftarrow S \cup \{k\}$
            else $S \leftarrow \{k\}$
        $k \leftarrow k + 1$
```

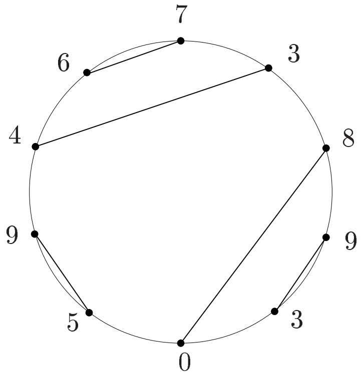

Figure 1: Illustration of HMatching.

```
return $|L|$
```

The action of the algorithm is illustrated in Figure 1.

In this example, $A[0] = 7, A[1] = 6, A[2] = 4, \ldots, A[9] = 3$. In the first step we explore $A[0]$ and add 0 to S. In the second step, we explore $A[1]$, and check if $A[0] + A[1] \pmod{10}$ is in L. Since it is not, the edge $(A[0], A[1])$ is added to $\mathcal{M}$, and since $A[0] + A[1] \ = 7 + 6 \ \equiv 3 \pmod{10}$, L becomes $\{3\}$, and S goes back to $\emptyset$. In the third step we explore $A[2]$, and since there is no stack, we add 2 to S (so that S becomes $\{2\}$, since it was empty) and move on to the fourth step, where we explore $A[3]$; since $A[2] + A[3] = 4 + 9 \equiv 3 \pmod{10}$ is already in L, we must now set $S = \{2, 3\}$. In the fifth step we explore $A[4] = 5$. Since 3 is the largest integer in S, we check if $A[3] + A[4] \equiv 4 \pmod{10}$ is in L. Since it is not, then we add $(A[3], A[4])$ to $\mathcal{M}$, 4 to L, and remove 3 from S. At the end, we obtain the matching shown, which happens to be perfect.

In words, this algorithm works as follows. By letting k increase from 0 to $n - 1$, we *explore* the labels $A[0], A[1], \ldots, A[n - 1]$ in the given order. Now at each step we have a set of matching edges (namely the current $\mathcal{M}$), whose set of labels is recorded as the set L, plus a *stack* of *unmatched* vertices (that is, not incident with an edge in $\mathcal{M}$), whose index set is S: the i–th vertex is unmatched iff $i \in S$. As we explore the next vertex label $A[k]$, we attempt to match it with the vertex i_0 in the stack such that i_0 is largest among all vertices in the stack. Note that this makes sense geometrically: if we manage to add this edge to $\mathcal{M}$ (that is, if $A[i_0] + A[k] \pmod{10} \notin L$), then the edge joining each vertex in the stack to each unexplored vertex does *not* cross any edge in $\mathcal{M}$. Loosely speaking, after we add a matching edge, each vertex in the stack (as well as each unexplored vertex, of course) still has a chance to be the endvertex of a matching edge.

We wrote code for this algorithm, including the generation of random labels, and ran it repeatedly for large values of n ($n = 10^4, 10^5, 10^6$). Its performance exceeded our best expectations. We consistently found a matching in question of size at least $0.999(n/2)$, thus extremely close to a perfect matching that contains $\lfloor n/2 \rfloor$ edges. In the next Section 3

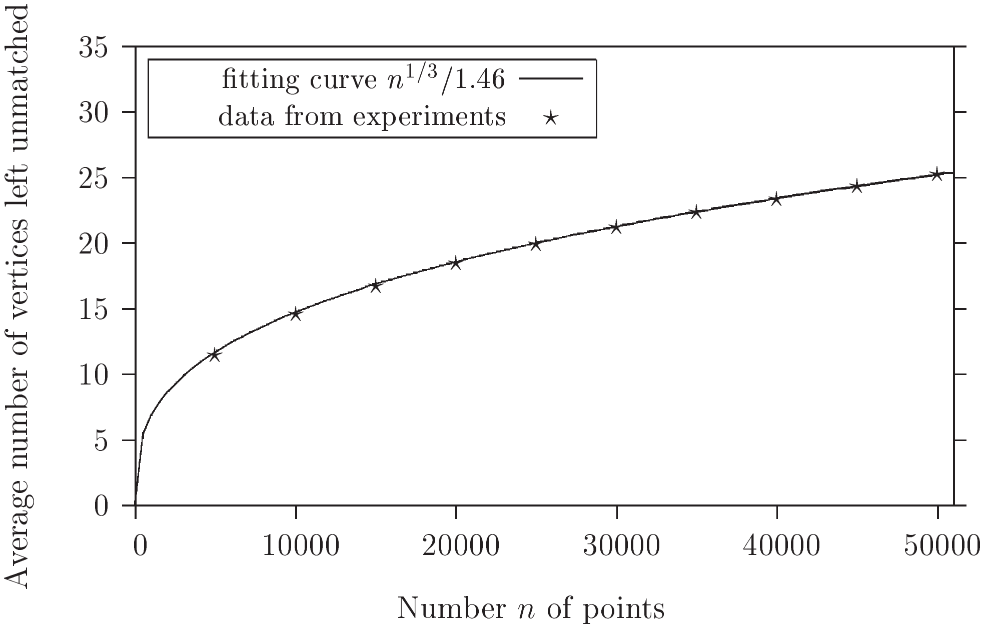

Figure 2: For each value of $n = 5000, 10000, 15000, \ldots, 50000$, we ran HMatching 10^6 times, each time on a randomly generated array $(A[0], A[1], A[2], \ldots, A[n-1]))$, and computed the average size of the set of vertices left unmatched. The results are shown in this figure, together with the fitting curve $n^{1/3}/1.46$ proposed by `Gnuplot`©.

we present our experimental results. Then in Section 4 we present a rigorous study which confirms our conjectures based on the experimental numerics.

3 Performance of HMatching: empirical results

There are two natural parameters to measure the performance of HMatching: (i) the expected size of the matching obtained by running HMatching, and (ii) the probability that HMatching delivers a perfect matching.

3.1 The empirical average size of matching

We wrote a `C++` program that generated each $A[i], i = 0, 1, \ldots, n-1$, independently at random from the uniform distribution on $\{0, 1, \ldots, n-1\}$, and then ran HMatching on $(A[0], A[1], \ldots, A[n-1])$ and returned the number of edges *left unmatched.*

We then ran this program 10^6 times for each of the following values of n: $5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000$, and 50000.

For each such n, we computed the average of the 10^6 experiments. Using `Gnuplot`©, we plotted the results and obtained a curve $n^{1/3}/1.46$ that fitted the data quite well. Both the results and the fitting curve are presented in Figure 2.

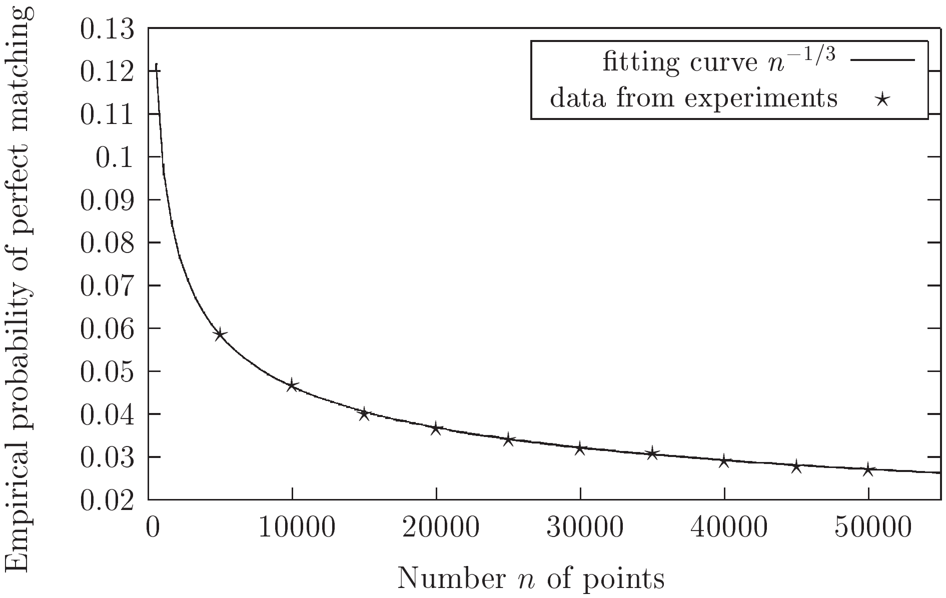

Figure 3: For each value of $n = 5000, 10000, \ldots, 50000$, we ran HMatching 10^6 times, each time on a randomly generated array $(A[0], A[1], A[2], \ldots, A[n-1])$, and computed the proportion of experiments for which HMatching delivered a perfect matching. The results are shown in this figure, together with the fitting curve $n^{1/3}$ proposed by `Gnuplot`©.

In view of the remarkably good fit given by the curve $n^{1/3}/1.46$, we conjecture that the expected number of vertices left unmatched is $\Theta(n^{1/3})$. An equivalent conjecture is that the expected size of the matching is $n/2 - \Theta(n^{1/3})$.

3.2 Empirical success probability.

Since on average the resulting matching turned out to be near perfect, we added a few lines of code to the program, to determine the empirical frequency of the problem instances when the matching was in fact perfect.

Again, we ran 10^6 experiments for each $n = 5000, 10000, \ldots, 50000$, and computed the proportion of experiments for which HMatching yielded a perfect matching. Using `Gnuplot`©, we plotted the results and got a fitting curve $n^{-1/3}$. As it can be checked in Figure 3, this curve seems to be a very good fit to the data obtained.

The data fit the curve $n^{-1/3}$ so well that we are led to the conjecture: the probability that HMatching delivers a perfect matching is of order $\Theta(n^{-1/3})$. It is tempting to state an even stronger conjecture: the probability that the resulting matching is perfect is asymptotic to $n^{-1/3}$. In the next section we prove a slightly weaker result, namely that this probability is between $c_1 n^{-1/3} \ln^{-1} n$ and $c_2 n^{-1/3} \ln n$. We also show that the likely size of the terminal matching is between $n/2 - c_3 n^{1/3} \ln n$ and $n/2 - c_4 n^{1/3} \ln^{-1} n$, which again is within the logarithmic factors from the conjectured formula $n/2 - \Theta(n^{2/3})$.

Consequently, on average, the number of the starting points for which the algorithm finds a perfect matching is of an empirical order $\Theta(n^{2/3})$, and of a provable order $\Omega(n^{2/3}\ln^{-1} n)$. This suggests the following

Conjecture 1 *W.h.p. there is a perfect (non–crossing, harmonic) matching, and it can be found by running* HMatching *n times, selecting each of the n points as a starting point.*

In our computer experiments, with n up to 10^5 and 10^6 problem instances, we always found a perfect matching by running the algorithm for sufficiently many starting points.

4 Analysis of HMatching

Recall that we are interested in analyzing the performance of HMatching when it is ran on an array $(A[0], A[1], \ldots, A[n-1])$ such that each $A[i]$ is an integer chosen uniformly at random, and independently of the other $A[j]$'s, from $\{0, 1, \ldots, n-1\}$.

4.1 The matching algorithm as a Markov Chain.

Consider the generic, k-th, step of the matching algorithm. Before this step the vertices $p_1, \ldots, p_{k-1}$ have been explored, and some of them have been matched. Let $\mathcal{M}$ be the current (non-crossing, harmonious) matching and S be the current set (stack) of all unmatched points whose labels have been explored. Then $2|\mathcal{M}| + |S| = k - 1$. Suppose first that $S \neq \emptyset$. Assume inductively that there are no triples (p_a, p_b, p_c), $a < b < c$, such that $(p_a, p_c) \in \mathcal{M}$ and $p_b \in S$. This condition means that no edge (p_a, p_b), such that $p_a \in S$ and $b > b^* = \max\{c : p_c \in S\}$, crosses an edge from $\mathcal{M}$. In particular, we can and do add to $\mathcal{M}$ the edge (p_{b^*}, p_k) if the label of this edge is not in L, the label set of the edges in $\mathcal{M}$, i. e. if $A[b^*] + A[k] \pmod n \ \notin L$. The last condition restricts the value $A[k]$ to a subset of $\{0, \ldots, n-1\}$ of cardinality $n - |L| = n - |\mathcal{M}|$. Since $A[k]$ is uniform on $\{0, \ldots, n-1\}$, and independent on $A[0], \ldots, A[k-1]$, the (conditional) probability that (p_{b^*}, p_k) is added to $\mathcal{M}$ in the k–th iteration step is $1 - |L|/n = 1 - |\mathcal{M}|/n$. In this case $\mathcal{M} + \{(p_{b^*}, p_k)\}$ and $S \setminus \{p_{b^*}\}$ are the next matching set and the next stack respectively. Alternatively, with the probability $|\mathcal{M}|/n$ the matching set remains the same, but the stack grows to $S \cup \{p_k\}$. If $S = \emptyset$, then the matching set $\mathcal{M}$ remains the same, and the next S is $\{p_k\}$. In all cases the new matching $\mathcal{M}$ and the new stack S meet the same non-crossing condition as the previous $\mathcal{M}$ and S. Clearly the sequence $\{\mathcal{M}_k, S_k\}_{k \le n}$, ($\mathcal{M}_0 = \emptyset$, $S_0 = \emptyset$), is a Markov chain. The chain terminates once $2|\mathcal{M}_k| + |S_k|$ reaches n, that is, when there are no unexplored points left. Remarkably, the transition probabilities and the termination rule depend only on $|\mathcal{M}_k|$. So there is a reduction of $\{\mathcal{M}_k, S_k\}$ to a much simpler Markov chain $\{m_k, s_k\}$ on the set of pairs (m, s), $m = |\mathcal{M}|$, $s = |S|$, with termination condition $2m_k + s_k = n$.

Here is the formal definition of the reduced Markov chain.

Markov Process 1 (MP$_1$) *Each state is a pair (m, s), where m and s are nonnegative integers, and $2m + s < n$, where n is a fixed integer given in advance. The initial state is $(0, 0)$. The transition rules are :*

If $s = 0$, then the next state is

$$(m, s + 1) = (m, 1).$$

If $s > 0$, then the next state is

$$\begin{aligned} (m + 1, s - 1), &\quad \textit{with probability } 1 - m/n, \\ (m, s + 1), &\quad \textit{with probability } m/n. \end{aligned}$$

4.2 The likely size of the terminal matching.

According to our reduction, to study the size of the terminal matching is equivalent to studying Z_n, the terminal value of m in the Markov chain MP$_1$.

Proposition 2 *(i) Given $a > 0$, set $\alpha = 2\sqrt{a(1 + a)}$.*

$$Pr(Z_n > n/2 - \alpha n^{1/3} \ln n) = 1 - O(n^{-a}). \tag{1}$$

(ii)

$$Pr(Z_n \le n/2 - n^{1/3} \ln^{-2} n) = 1 - O(\ln^{-1} n). \tag{2}$$

(iii) Let $P_n = Pr(Z_n = n/2)$, n even, and $P_n = Pr(Z_n = (n - 1)/2)$, n odd. Then, for some constants $\alpha, \beta > 0$,

$$\alpha n^{-1/3} \ln^{-1/2} n \le P_n \le \beta n^{-1/3} \ln n. \tag{3}$$

For the proof we need the following statement.

LEMMA 1 *Let $a > 0$. With probability $1 - O(n^{-a})$, there exists k such that*

$$m_k \in (n/2 - (1 + a)n^{2/3} \ln n,\ n/2 - 0.5(1 + a)n^{2/3} \ln n), \quad s_k = 0,$$

with $\Theta(n^{2/3} \ln n)$ points remaining to be explored.

Proof of Lemma 1. Given $m < n/2$, let $T_m = \min\{k : m_k = m\}$ and set $T_m = n$, if no such k exists. Introduce $H_m = s_{T_m}$, the stack size at this moment. By the definition of MP$_1$, for $j < k$ and $s_j > 0$, the conditional probability of the transition $(m_j, s_j) \to (m_{j+1}, s_{j+1}) = (m_j, s_j + 1)$, which leads to an increase of the stack by 1, is m/n at most. And the alternative transition leads to the stack size $s_j - 1$. For $s_j = 0$, we have $s_{j+1} = 1$. These observations imply that H_m is stochastically dominated by W_m, the maximum of the simple asymmetric random walk $\{\xi_j\}_{j\le n}$ on $\{0,1,2,\ldots\}$, defined as follows: $\xi_0=0$,

$$\begin{aligned}&\Pr(\xi_{j+1}=\xi_j+1\,|\,\xi_j)=p:=m/n,\quad (\xi_j\ge 1),\\ &\Pr(\xi_{j+1}=\xi_j-1\,|\,\xi_j)=q:=1-m/n,\quad (\xi_j\ge 1),\\ &\Pr(\xi_{j+1}=1\,|\,\xi_j=0)=1.\end{aligned}$$

Furthermore, for each integer $w>0$, $\Pr(W_m>w)\le n\Pr(\mathcal{W}_m>w)$, where $\mathcal{W}_m$ is the maximum of ξ_j for j between 0 and the first moment $t>0$ when $\xi_t=0$. Using the classic gambler's ruin formula, we have

$$\Pr(\mathcal{W}_m>w)=\frac{q/p-1}{(q/p)^{w+1}-1}\le (p/q)^w.$$

Then, introducing $m_i=\frac{n}{2}-[a_in^{2/3}\ln n]$ and $p_i=m_i/n$, $i=1,2$, with $a_2=a_1/2$, we have

$$\begin{aligned}&\Pr\big(W_{m_i}>n^{1/3}\big)\le 2n\left(\frac{m_2}{n-m_2}\right)^{n^{1/3}}\\ &<3n\big(1-4a_2n^{-1/3}\ln n\big)^{n^{1/3}}<3n\exp(-4a_2\ln n)=\frac{3}{n^{2a_1-1}}\ \to\ 0,\end{aligned}$$

provided $a_1>1/2$.

Now, since $2m_k+s_k=k$ at each step, we have

$$m=\frac{T_m-H_m}{2}\ge\frac{T_m-W_m}{2}.$$

Applying this to $m=m_1,m_2$, we see that

$$\Pr\left\{\bigcap_{i=1}^{2}\big(2m_i\le T_{m_i}\le 2m_i+n^{1/3}\text{ and }H_{m_i}\le n^{1/3}\big)\right\}\ge 1-O(n^{-a}),\quad a=2a_1-1.$$

Therefore

$$\Pr\left\{(T_{m_2}-T_{m_1}=a_1n^{2/3}\ln n+O(n^{1/3}))\cap(H_{m_1}\le n^{1/3})\right\}\ge 1-O(n^{-a}).$$

Denote the event in this bound by A. Let

$$B=A\cap\{s_k\text{ becomes zero at some }k\in[T_{m_1},T_{m_2}]\}.$$

We want to show that $\Pr(A\setminus B)\le n^{-b}$, $\forall\, b>0$, for n large enough. Let $t_1\in[0,n-1]$. Suppose that $m_{t_1}\le m_2$, and $0<s_{t_1}\le n^{1/3}$. These conditions certainly hold if $t_1=T_{m_1}$. Let $\mathcal{T}=\mathcal{T}(t_1)$ be the first $t>t_1$ such that either $m_t=m_2$, or $s_t=0$. As before, $\{s_t\}_{t<\mathcal{T}}$ is dominated by the asymmetric walk $\{\xi_j\}_{j\ge t_1}$, $\xi_{t_1}=\lfloor n^{1/3}\rfloor$, with $p=m_2/n$. Therefore $\mathcal{T}-t_1$ is dominated by $X_{p,\lfloor n^{1/3}\rfloor}$, where $X_{p,s}$ is the first time the random walk hits 0, if $\xi_0=s$. Since (see Proposition 8 in the Appendix)

$$\Pr(X_{p,s}\ge r)\le\frac{(4pq)^{-r/2}}{(2p)^s},$$

it follows that

$$\Pr\left(X_{\frac{m_2}{n},\lfloor n^{1/3}\rfloor}\ge\lfloor n^{2/3}\rfloor\right)\ \le\ \frac{\left(1-\frac{4\lfloor a_2n^{2/3}\ln n\rfloor}{n^2}\right)^{\lfloor n^{2/3}\rfloor}}{\left(1-\frac{2\lfloor a_2n^{2/3}\ln n\rfloor}{n}\right)^{\lfloor n^{1/3}\rfloor}}\ \le\ \exp(-a_2\ln^2 n).$$

Therefore $\mathcal{T}(t_1)-t_1\le n^{2/3}$ quite surely (q.s. in short), i.e. with probability $1-n^{-b}$, for every $b>0$, uniformly for all t_1. Thus $\mathcal{T}(T_{m_1})-T_{m_1}\le n^{2/3}$ q.s. as well. Since $T_{m_2}-T_{m_1}$ is of order $n^{2/3}\ln n\gg n^{2/3}$ on A, we conclude that indeed $\Pr(A\setminus B)\le n^{-b}$, for every $b>0$. So the Markov process $\{m_k,s_k\}$ reaches a state $(m_0,0)$, where $n/2-m_0\in(0.5a_1n^{2/3}\ln n,\ a_1n^{2/3}\ln n)$, with probability $1-O(n^{-a})$, $a=2a_1-1$. ∎

Proof of Proposition 2 (i) Let T be the first k such that

$$m_k\in(n/2-(1+a)n^{2/3}\log n,\,n/2-0.5(1+a)n^{2/3}\log n),\quad s_k=0.$$

By Lemma 1, T is well defined with probability $1-O(n^{-a})$. Let ℓ be the number of the remaining unexplored points after T steps; clearly

$$(1+a)n^{2/3}\ln n\le\ell\le2(1+a)n^{2/3}\ln n.$$

The additional increase of m_k during the remaining $n-T$ steps is $(\ell-s_n)/2$, where s_n is the terminal stack size. So $Z_n=m_n$ is given by

$$Z_n=\frac{n-\ell}{2}+\frac{\ell-s_n}{2}=\frac{n}{2}-0.5s_n.$$

Thus we need to show that w.h.p. $s_n=O(n^{1/3}\log n)$. Since $m_k\le n/2$ for all k, s_n is dominated by ξ_ℓ, where $\{\xi_j\}$ is the simple symmetric random walk with $p=q=1/2$, and $\xi_0=0$. We need to find a likely upper bound for ξ_ℓ. First of all, for each integer $x\ge0$,

$$\Pr(\xi_\ell=x)=\sum_{2t+\mu=\ell}\mathcal{P}_t\mathcal{Q}_\mu(x);\tag{4}$$

here $\mathcal{P}_t=\Pr(\xi_{2t}=0)$, the probability that the walk returns to 0 after $2t$ steps; $\mathcal{Q}_\mu(0)=\delta_{\mu,0}$, and $\mathcal{Q}_\mu(x)$, $x>0$, is the probability that the walk, that starts at 0, reaches x after μ steps without ever returning to 0. We will need the full strength of this formula later, but for now we are content with its weak corollary, namely

$$\Pr(\xi_\ell=x)\le\sum_{\substack{\mu,t\ge0\\2t+\mu=\ell}}\mathcal{Q}_\mu(x).\tag{5}$$

As for $\mathcal{Q}_\mu(x)$, recall that, by the ballot theorem, the total number of ways to reach the point x from the point 0 by making μ (± 1)-moves, without returning to 0, is

$$\frac{x}{\mu}\binom{\mu}{(\mu+x)/2}, \quad \mu \geq x,$$

$(\mu+x)/2$ being the total number of right moves. Therefore, for the (p,q)-simple walk,

$$\mathcal{Q}_\mu(x) := \frac{x}{\mu}\binom{\mu}{(\mu+x)/2} p^{(\mu+x)/2-1} q^{(\mu-x)/2};$$

(the probability of the first move, from 0 to 1, is 1, each of the other $\mu - 1$ moves has probability p.) Using Stirling's formula and $4pq \leq 1$, we obtain a simple estimate

$$\mathcal{Q}_\mu(x) \leq c_0 x \frac{\exp(-x^2/2\mu)}{\sqrt{\mu(\mu^2 - x^2 + \mu)}}, \quad x > 0 \tag{6}$$

where c_0 is some constant. (We will continue to use c's for various absolute constants.) Combining (5) and (6), we have

$$\Pr(\xi_\ell = x) \;\leq\; c_0 x e^{-x^2/2\ell} \sum_{x\leq\mu\leq\ell} \frac{1}{\sqrt{\mu(\mu^2-x^2+\mu)}} \;\leq\; c_1 x e^{-x^2/2\ell}.$$

(That the last sum is uniformly bounded follows from considering separately $\mu \geq 2x$ and $x \leq \mu \leq 2x$.) Then

$$\Pr(\xi_\ell \geq \alpha n^{1/3} \ln n) \;\leq\; c_1 \sum_{x \geq \alpha n^{1/3} \ln n} x e^{-x^2/2\ell} \;\leq\; c_2 \exp\left(-\frac{\alpha^2 n^{2/3} \ln^2 n}{4(1+a) n^{2/3} \ln n}\right) \;=\; c_2 n^{-a}, \tag{7}$$

as $\ell \leq 2(1+a) n^{2/3} \ln n$, and $\alpha = 2\sqrt{a(1+a)}$. ∎

Proof of Proposition 2 (ii) As in the proof of part 1,

$$Z_n = \frac{n}{2} - 0.5 s_n,$$

so we need to show that w.h.p. $s_n \geq \nu_n := 2n^{1/3} \ln^{-2} n$. Clearly s_n stochastically dominates ξ_ℓ for the (p,q)-walk, where

$$p \;=\; \frac{m_T}{n} = \frac{1}{2} - \frac{\ell}{n} \quad \text{and} \quad \ell \;=\; n - 2m_T \in [(1+a)n^{2/3} \ln n,\; 2(1+a) n^{2/3} \ln n].$$

Thus

$$\Pr(s_n \leq \nu_n) \leq \Pr(\xi_\ell \leq \nu_n) = \sum_{x \leq \nu_n} \Pr(\xi_\ell = x),$$

with $\Pr(\xi_\ell = x)$ given by (4). This time we need a sharp bound for $\mathcal{P}_t$, which is

$$\mathcal{P}_t \leq c((1-2p) + (t+1)^{-1/2}) = c\left(\frac{\ell}{n} + t^{-1/2}\right), \tag{8}$$

see Proposition 9. For $t \in [\ell/2, \ell]$, the first summand dominates since $\ell^{3/2} \gg n$, and the bound simplifies to $\mathcal{P}_t \leq 2c(\ell/n)$. Break the sum in (4) into two parts, $\mu \geq \ell/2$ and $\mu < \ell/2$. Since $x \leq \nu_n \ll \ell$, it follows from (6) and (8) that, for $x > 0$,

$$\begin{aligned}\sum_{\substack{2t+\mu=\ell\\ \mu\geq\ell/2}} \mathcal{P}_t \mathcal{Q}_\mu(x) &\leq c'x \left[\sum_{\mu\geq\ell/2} \mu^{-3/2}\big(\ell/n + (\ell-\mu+1)^{-1/2}\big)\right]\\ &\leq c''x\big((\ell/n)\ell^{-1/2} + \ell^{-1}\big) = O\big(x\ell^{1/2}/n\big),\end{aligned}$$

as $\ell^{3/2} \gg n$. Therefore

$$\sum_{0<x\leq\nu_n} \sum_{\substack{2t+\mu=\ell\\ \mu\geq\ell/2}} \mathcal{P}_t \mathcal{Q}_\mu(x) = O(\nu_n^2 \ell^{1/2} n^{-1}) = O(\ln^{-3/2} n). \tag{9}$$

Let $\mu \leq \ell/2$ now. Since $2t + \mu = \ell$, it follows that $t \geq \ell/4$, and so $\mathcal{P}_t = O(\ell/n)$. Then, using (6), we obtain

$$\sum_{\substack{t+\mu=\ell\\ \mu\leq\ell/2}} \mathcal{P}_t \mathcal{Q}_\mu(x) \leq \hat{c}\ell n^{-1} x \left(\int_x^\infty \frac{e^{-x^2/2y}}{\sqrt{y(y^2-x^2)}}\, dy\right). \tag{10}$$

Substituting $y = x/z$, we transform the last integral into

$$x^{-1/2} \int_0^1 \frac{e^{-xz/2}}{\sqrt{z(1-z^2)}}\, dz = \sqrt{\frac{2}{x}}\,(J_1 + J_2),$$

with J_1, J_2 corresponding to integration over $[0, 1/2]$ and $[1/2, 1]$, respectively. Then, substituting $w = xz/2$,

$$J_1 \leq \frac{2}{\sqrt{3}} \int_0^{1/2} z^{-1/2} e^{-xz/2}\, dz \leq \frac{2}{\sqrt{3}} x^{-1/2} \int_0^\infty w^{-1/2} e^{-w}\, dw = \hat{c}_1 x^{-1/2},$$

and

$$J_2 \leq e^{-x/4} \int_{1/2}^1 \frac{dz}{\sqrt{z(1-z^2)}} = \hat{c}_2 e^{-x/4}.$$

Therefore the bound (10) becomes

$$\sum_{\substack{2t+\mu=\ell\\ \mu\leq\ell/2}} \mathcal{P}_t \mathcal{Q}_\mu(x) = O(\ell n^{-1} x (x^{-1/2})^2) = O(\ell/n), \quad x > 0.$$

Consequently

$$\sum_{0<x\leq\nu_n} \sum_{\substack{t+\mu=\ell\\ \mu\leq\ell/2}} \mathcal{P}_t \mathcal{Q}_\mu(x) = O(\ell n^{-1} \nu_n) = O(\ln^{-1} n). \tag{11}$$

Combining (9) and (11), we obtain

$$\sum_{0<x\le\nu}\sum_{2t+\mu=\ell}\mathcal{P}_t\mathcal{Q}_\mu(x)=O(\ln^{-1}n).$$

Finally

$$\sum 2t+\mu=\ell\mathcal{P}_t\mathcal{Q}_\mu(0) \;=\; \mathcal{P}_{\ell/2} \;=\; O(\ell/n) \;=\; O(n^{-1/3}\ln n).$$

So

$$\Pr(\xi_\ell\le\nu_n) \;=\; \sum_{0\le x\le\nu_n}\sum_{2t+\mu=\ell}\mathcal{P}_t\mathcal{Q}_\mu(x) \;=\; O(n^{-1/3}\ln n)+O(\ln^{-1}n) \;=\; O(\ln^{-1}n).$$

$\blacksquare$

Since $Z_n=n/2-0.5s_n$, and s_n dominates ξ_ℓ, the statement follows.

Proof of Proposition 2 (iii) First of all, for n even, $Z_n=n/2$ iff $s_n=0$, and, for n odd, $Z_n=(n-1)/2$ iff $s_{n-1}=0$. Consider, for instance, even n. We know that, conditioned on the event in Lemma (call it $\mathcal{A}$), s_n is dominated by $\xi_\ell(1/2)$ of the walk $(\{\xi_j\}_{j\le\ell})$ with $p=1/2$, and dominates ξ_ℓ of the walk with $p=p_n:=1/2-\ell/n$. Then, using (8),

$$\Pr(s_n=0\,|\,\mathcal{A}) \;\le\; \Pr(\xi_r(p_n)=0)\big|_{r=\ell} \;=\; O((1-2p_n)+\ell^{-1/2}) \;=\; O(\ell/n) \;=\; O(n^{-1/3}\ln n).$$

On the other hand, again using (8),

$$\Pr(s_n=0\,|\,\mathcal{A}) \;\ge\; \Pr(\xi_r(1/2))\big|_{r=\ell} \;=\; \Omega(\ell^{-1/2}) \;=\; \Omega(n^{-1/3}\ln^{-1/2}n).$$

Since $\Pr(\mathcal{A}^c)=O(n^{-a})$, picking $a>1/3$ we conclude that unconditionally

$$\alpha n^{-1/3}\ln^{-1/2}n\le\Pr(s_n=0)\le\beta n^{-1/3}\ln n,$$

for some absolute constants $\alpha,\beta>0$. The case n odd is similar.

This completes the proof of the proposition. $\blacksquare$

5 Random permutation labeling.

The random labeling we have studied very likely assigns the same labels to different points. (Indeed, the probability that no two points have the same label is $n!/n^n\ll 2^{-n}$.) If we consider only distinct labels, then it is natural to assume that the labels of n points form the uniformly random permutation of $\{0,\dots,n-1\}$. We strongly believe that our results on matchings and paths continue to hold for this uniform permutation labeling, and the computer experiments provide an ample evidence supporting this belief. However, a rigorous proof of such an "invariance" is quite problematic. We model the work of our algorithms as the processes, in which at every step we explore the label of a point for the first time. So, for the independent labels, this label is conditionally uniform on $\{0,\dots,n-1\}$, while for the random permutation labeling it is conditionally uniform on the subset of labels not yet seen. This complication makes it necessary to keep track of the labels encountered so far, thus invalidating usage of the relatively simple Markov chain $\{m_k, s_k\}$. Whether the corresponding Markov chain can be rigorously analyzed is, in our opinion, a challenging problem.

Here is a version of the matching problem for which we can prove the asymptotic equivalence of the two labelings. Let $P = \{p_0, p_1, \dots, p_{n-1}\}$ and $Q = \{q_0, q_1, \dots, q_{n-1}\}$ be such that the points of $P \cup Q$ lie on a circle, in the cyclic order $p_0, p_1, \dots, p_{n-1}, q_0, q_1, \dots, q_{n-1}$. We consider *parallel* matchings between P and Q, that is, matchings consisting of the edges (p_i, q_i) only. Clearly the maximum (harmonic) matching size equals D_n, the total number of distinct labels $\ell(p_i, q_i)$ $(\equiv (\ell(p_i) + \ell(q_i) \pmod{n}))$. Suppose that $\ell(p_i) = i$, $0 \le i \le n-1$, and that the labels of the points in Q are either independent, uniform on $\{0, \dots, n-1\}$, or are the elements of the uniformly random permutation of $\{0, \dots, n-1\}$. Then, for each of the labelings, $D_n/n \to 1 - e^{-1}$ in probability. The proofs are based on evaluation of the two first order moments of D_n, but the computations for the random permutation case are more involved.

6 Non–crossing harmonic paths

We now turn to Question 2, which is: how many edges are there typically in a maximum size harmonic non–crossing path in G_S?

6.1 Roster of algorithms.

Again, there is given a collection $S = \{p_0, p_1, \dots, p_{n-1}\}$ of points in convex position, and the labels $A[i] = \ell(p_i)$ are independent, uniform on $\{0, \dots, n-1\}$, while each edge (p_i, p_j) is assigned a label $\ell(p_i, p_j) :\equiv (\ell(p_i) + \ell(p_j)) \pmod{n}$.

We seek algorithms that take as input an array $(A[0], A[1], \dots, A[n-1])$, and deliver a provably long path on G_S which is both non–crossing and harmonic.

We studied the following algorithms.

(i) GPath1. It is relatively simple to study, but the resulting path is disappointingly short on average.

(ii) GPath2. Besides being quite natural, this algorithm typically delivers an impressively long path. In fact, in our experiments it outperformed all other variants of path algorithms. However we could not analyze its performance rigorously.

(iii) HPath. Its empirical performance puts it above GPath1 and below GPath2. Crucially, HPath is amenable to a rigorous analysis, which confirms its superiority over GPath1.

6.2 GPath1.

GPath1 works as follows. Start with p_0. Recursively, given a current path $p_0 = p_{i_0} \to p_{i_1} \to \dots \to p_{i_k}$, enlarge it by adding the first vertex from $\{p_i : i > i_k\}$ such that the resulting path $p_0 = p_{i_0} \to \dots \to p_{i_k} \to p_{i_{k+1}}$ remains harmonic. This is equivalent to the condition $A[i] \notin \{A_1, \dots, A_k\}$, where $A_1, \dots, A_k$ are determined by the labels of the edges $(p_{i_j}, p_{i_{j+1}})$, $0 \le j \le k-1$. The new path remains non–crossing automatically since $i_k > i_{k-1}$. Continue, until no such enlargement is possible. Obviously, for each k, no vertex p_i with $i > i_k$ has been tested as a candidate for joining the path until it has grown to length k.

Recall that $A[0], \dots, A[n-1]$ are independent, each uniformly distributed on $\{0, \dots, n-1\}$. Then, given the labels $A[j]$, $j \le i_k$, the labels $A[i]$, $i > i_k$, remain mutually independent, and uniform. So the events $A[i] \notin \{A_1, \dots, A_k\}$, $i > i_k$, are conditionally independent, each of the conditional probability $1 - k/n$. It follows then that the length of the terminal path has the same distribution as D_n, the number we encountered studying the largest size of the parallel matching. Thus the likely number of edges in the terminal path is asymptotic, in probability, to $(1 - 1/e)n \approx 0.632n$.

6.3 GPath2.

Unlike GPath1, in each step of GPath2 there are two possible types of a point to be added to the current path. At the end of the k–th step we have: (i) the current (non–crossing, harmonic) path

$$P_k = \{p_{i_0} \to p_{i_1} \to \dots \to p_{i_\ell}\}, \quad \ell = \ell(k) \le k,\, p_{i_0} = p_0,\, p_{i_1} = p_{n-1};$$

(ii) the set

$$D_k = \{p_{L_k}, p_{L_k - 1}, \dots, p_1\,;\, p_0, p_{n-1}, p_{n-2}, \dots, p_{R_k+1}, p_{R_k}\}, \quad L_k < R_k,$$

of *dead* points, never to be used in future for extending a path. In particular, $P_k \subseteq D_k$. In addition, p_{i_ℓ}, the newest vertex of the path P_k, is either p_{L_k} or p_{R_k}. Let $G_k := S - D_k$ denote the set of *game* points, that is, the points that still can be added to the path P_k.

For instance, at the end of the first step of GPath2 the path P_1 is $\{p_0 \to p_{n-1}\} = \{p_{i_0} \to p_{i_1}\}$, $L_1 = 0$ and $R_1 = n-1$, $D_1 = \{p_0, p_{n-1}\}$. In general, D_k may well contain the points other than those from P_k.

Clearly any vertex from G_k can be added to P_k without causing new edge cross any edge of P_k. So our only concern is that a new edge must have a label different from the labels of all edges in the path P_k. Suppose, for instance, that $p_{i_\ell} = p_{R_k}$. First we test p_{L_k+1}, the point that follows D_k in the counterclockwise direction. If it fails the test, (i. e. if the label of (p_{L_k+1}, p_{i_ℓ}) has been encountered earlier), then we test p_{R_k-1}, the point that follows D_k in the clockwise direction. We keep testing new points in this alternating fashion until we find a point $p_{i_{\ell+1}}$ that can be joined to p_{i_ℓ}, to extend the current path P_k. If $p_{i_{\ell+1}} = p_{L_k+t}$, $(t \ge 1)$, then all the points P_{L_k+s}, $(0 < s < t)$, "die", so that

$$D_{k+1} = D_k \cup \{p_{L_k+s} : 0 < s < t\}.$$

If $p_{i_{\ell+1}} = p_{R_k - t}$, $(t \geq 1)$, then

$$D_{k+1} = D_k \cup \{p_{R_k - s} :\ 0 < s < t\}.$$

If no such point is found, then the process stops. GPath2 is illustrated in Figure 6.3.

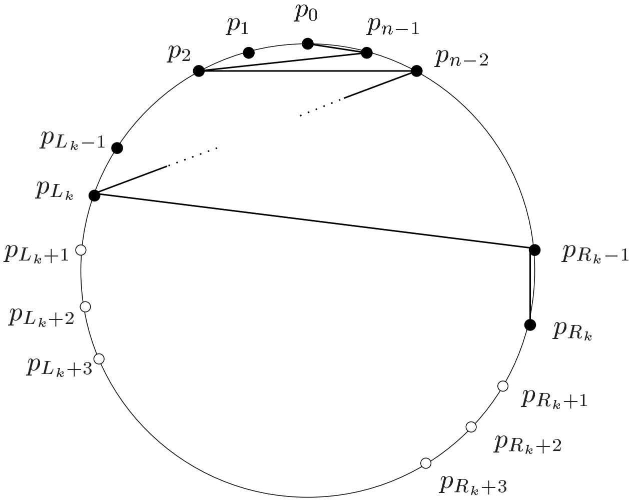

Figure 4: Illustration of GPath2.

Unlike the algorithms we have discussed, G_k (the set of active (game) vertices) may contain, in addition to fresh vertices, some vertices whose labels had been tested in the previous steps. This diversity makes it hard to bound from below the (conditional) probability that a game vertex passes the test, and can be chosen as the next extension of the current path.

Conjecture 3 *The likely number of edges in a path delivered by* GPath2 *is asymptotic to*

$$\frac{e^2 - 1}{e^2 + 1} n \approx 0.761n.$$

Our extensive computer experiments support the estimate $0.761n$ for the average path length. The number $(e^2-1)/(e^2+1)$ comes from the following simple-minded model of the algorithm. We assume that the label of a vertex being tested is generated anew, uniformly at random on $\{0, \dots, n-1\}$, and independently of all other explored labels, including the old label of the vertex in question if it had been tested before. This assumption erases the difference between the old and the new vertices in G_k, and the probability that a vertex in G_k can be added to the current path P_k is simply $1 - |P_k|/n$. Let $\pi_k = |P_k|$, and let d_k denote the total number of dead vertices not in P_k. Then $\{\pi_k, d_k\}$ is a Markov chain such that: $(\pi_0, d_0) = (0, 0)$, and

$$(\pi_{k+1}, d_{k+1}) = (\pi_k + 1, d_k) \text{ with probability } 1 - (\pi_k/n)^2,$$

$$(\pi_{k+1}, d_{k+1}) = (\pi_k, d_k + 1) \text{ with probability } (\pi_k/n)^2.$$

The process terminates when $\pi_k + d_k = n - 1$.
A coupon-collector type of argument shows that the likely length of the terminal path is asymptotic to cn, where c is the solution of the equation

$$\int_0^c \frac{1}{1-x^2}dx = 1,$$

or $c = (e^2 - 1)/(e^2 + 1) \approx 0.761$. The challenge is to show that whp the work of the actual algorithm is asymptotically close to this Markov process.

6.4 HPath: the algorithm

In this section, lowering our sights, we describe an algorithm (HPath) which on average performs better than GPath1, but falls short of the conjectured performance of GPath2. Unlike GPath1, in HPath some of the vertices that had failed the label test are tested again, and this modification typically leads to fewer wasted vertices.

As in GPath2, at the end of the k-th step we have the current path $P_k = \{p_{i_0} \to p_{i_1} \to \dots \to p_{i_\ell}\}$, $\ell = \ell(k) \le k$, $(p_{i_0} = p_0,\, p_{i_1} = p_{n-1})$, and the set $D_k \supseteq P_k$ of the "dead" vertices. D_k is an interval, whose endpoints are in P_k. One of the endpoints is p_{i_ℓ}. The set D_k^c of the "live" vertices contains an interval F_k of the "fresh" vertices, i. e. vertices whose labels have not been tested yet. In general, D_k and F_k are separated by two, left and right, intervals consisting of vertices already tested. Denote these intervals T_k^L and T_k^R and let $T_k = T_k^L \cup T_k^R$. At most one vertex u in T_k may still be alive, in which case its label $A[u]$ is different from the label of p_{i_ℓ}, *and* the path extension $p_{i_\ell} \to u$ is unfeasible. The remaining vertices in T_k are dead.

Case 1. T_k consists of dead vertices only. Picking the left endpoint q of F_k, we check whether the label of the edge (p_{i_ℓ}, q) is different from the labels on the edges in P_k, so that (p_{i_ℓ}, q) can be added to P_k. If it can, then we set $P_{k+1} = \{p_{i_0} \to \dots \to p_{i_\ell} \to q\}$, so that $\ell(k+1) = \ell(k) + 1$ and $p_{i_{\ell(k+1)}} = q$. Furthermore $D_{k+1} := D_k \cup T_k^L$, $T_{k+1}^L := \emptyset$, $T_{k+1}^R := T_k^R$, $F_{k+1} := F_k \setminus \{q\}$. T_{k+1} consists of dead vertices only. If (p_{i_k}, q) cannot be added to P_k and the labels of p_{i_ℓ}, q are the same then q is declared dead. If the labels of p_{i_ℓ} and q are distinct, then q is declared alive. In either case, $T_{k+1}^L := T_k^L \cup \{q\}$, $T_{k+1}^R := T_k^R$, $F_{k+1} := F_k \setminus \{q\}$, $D_{k+1} := D_k$, $P_{k+1} := P_k$. Clearly at most one vertex u in T_{k+1} is alive, in which case: (i) $A[u] \ne A[p_{i_{\ell(k+1)}}]$; (ii) the path extension $p_{i_{\ell(k+1)}} \to u$ is unfeasible.

Case 2. T_k contains exactly one vertex (call it u) still alive. Suppose, say, that $u \in T_k^L$. Consider the right endpoint v of the fresh interval F_k such that F_k is sandwiched between u and v.

Case 2(a). If the path extension $p_{i_\ell} \to v$ is unfeasible, then v is declared dead, and $T_{k+1}^L := T_k^L$, $T_{k+1}^R := T_k^R \cup \{v\}$, $F_{k+1} := F_k \setminus \{v\}$, $D_{k+1} := D_k$, $P_{k+1} := P_k$. u is a sole alive vertex in T_{k+1}, its label meeting the conditions (i), (ii), (see Case 1).

Case 2(b). Suppose $p_{i_\ell} \to v$ can be used for extending the path P_k. We check then whether the 2-edge extension $p_{i_\ell} \to v \to u$ is usable as well.

In case "no" we set $P_{k+1} := \{p_{i_0} \to \ldots \to p_{i_\ell} \to v\}$, $T^L_{k+1} := T^L_k$, $T^R_{k+1} := \emptyset$, $F_{k+1} := F_k \setminus \{v\}$, $D_{k+1} := D_k \cup T^R_k \cup \{v\}$. Note that T_{k+1} still contains u, and the label of u is different from the label of $v = p_{i_{\ell(k+1)}}$. Otherwise, like $p_{i_{\ell(k)}} \to v$, $p_{i_{\ell(k)}} \to u$ would have also been a feasible extension of P_k, which contradicts the definition of u, the sole alive vertex in T_k. And, of course, $p_{i_{\ell(k+1)}} \to u = v \to u$ is not a feasible path extension. Thus u is the sole alive vertex in T_{k+1}, and the conditions (i), (ii) are met again.

In case "yes" we set $P_{k+1} := \{p_{i_0} \to p_{i_1} \to \ldots \to p_{i_\ell} \to v \to u\}$, so $\ell(k+1) = \ell(k) + 2$, and $p_{i_{\ell(k+1)}} = u$. We set $T^R_{k+1} := \emptyset$, $F_{k+1} := F_k \setminus \{v\}$, $D_{k+1} := D_k \cup T^R_k \cup \{v\} \cup A$ and $T^L_{k+1} := T^L_k \setminus [p_{i_\ell}, u]$. Here T_{k+1} consists of dead vertices only.

The process stops when F, the set of fresh vertices, becomes empty.

7 Analysis of HPath

The main result in this section is a lower bound for the expected number of edges in the path delivered by HPath.

Theorem 4 *Let*

$$\alpha = -\ln 2 + \frac{3}{2\sqrt{5}} \ln\left(\frac{(\sqrt{5}+2)(\sqrt{5}-1)}{(\sqrt{5}-2)(\sqrt{5}+1)}\right) \approx 0.598.$$

Then the expected number of edges in a path obtained by the action of HPath *is at least*

$$\left(1 - \frac{e^{\alpha-1}}{2}\right) n \approx 0.665n.$$

Thus w.h.p. HPath outperforms GPath1.

Proof. We break the analysis into two parts. First we obtain a probabilistic upper bound for the number of vertices it takes to build a path of length $n/2$. Second, we use a balls-into-boxes argument to bound from below the expected number of edges added to the path during the remaining steps.

Lemma 5 *Let η_n denote the random number of vertices tested by the algorithm till the current path length reaches $n/2$. Then, for each $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left(\frac{\eta_n}{n} \le (1+\varepsilon)\alpha\right) = 1. \tag{12}$$

Proof. (I) After the k-th step, we have the current path $P = P_k$, the set of dead vertices $D = D_k \supset P_k$, and the set $T = T_k$ of other vertices, already tested, that separates D from $F = F_k$, the set of fresh vertices. A vertex $u \in T$ is singled out as an only vertex in T, still alive, if it is present. To complete the description of the current state we need to list the labels $A[i]$ of vertices i from P and the label of a still alive vertex $u \in T$, if it exists; in that case $A[u] \neq A[p_{end}]$, where p_{end} being the endvertex of P, and u cannot be used to extend P. Let $\mathcal{S}$ be the resulting state description. It can be seen that the sequence $\{\mathcal{S}_k\}$ is a Markov chain. As in the case of the matching algorithm, it is possible to determine a much simpler Markov chain dominated by $\{\mathcal{S}_k\}$. Let ℓ be the length of the current path P. Let $\sigma \in \{0, 1\}$ be an indicator of the event $\{T$ contains an alive vertex $u\}$. Denote by ℓ', σ' the parameter of the next state $\mathcal{S}'$, we have:

$$P[(\ell', \sigma') \ = \ (\ell + 1, \sigma) \mid \mathcal{S}] \ = \ 1 - \ell/n, \tag{13}$$
$$P[(\ell', \sigma') \ = \ (\ell, \sigma + 1) \mid \mathcal{S}] \ \geq \ (\ell - 1)/n, \tag{14}$$
$$P[(\ell', \sigma') \ = \ (\ell, \sigma) \mid \mathcal{S}] \ \leq \ 1/n, \tag{15}$$

if $\sigma = 0$, and

$$P[(\ell', \sigma') \ = \ (\ell, \sigma) \mid \mathcal{S}] \ = \ \ell/n \tag{16}$$
$$P[(\ell', \sigma') \ = \ (\ell + 2, \sigma - 1) \mid \mathcal{S}] \ \geq \ 1 - 2\ell/n \tag{17}$$
$$P[(\ell', \sigma') \ = \ (\ell + 1, \sigma) \mid \mathcal{S}] \ \leq \ \ell/n, \tag{18}$$

if $\sigma = 1$.

Let us prove (13)–(18). Suppose $\sigma = 0$. The relation (13) follows from the observation that a fresh vertex v can be added to P iff its label is not equal to one of ℓ "excluded" values, determined by the edge labels of the path P *and* and $A[p_{\text{end}}]$, i. e. iff $A[v] \notin \text{Ex}(P, p_{\text{end}})$, $|\text{Ex}(P, p_{\text{end}})| = \ell$. Then the sum of two other conditional probabilities is ℓ/n, and the third probability is at most $1/n$, the probability that the fresh vertex has the same label as $A[p_{end}]$ (and could not be added to the path).

Suppose now that $\sigma = 1$. Then (16) holds, analogously to (13), and thus the sum of two other probabilities is $1 - \ell/n$. So we need to prove (17) only. If a fresh vertex $v \in F$ cannot be added to P then $A[v] \in \text{Ex}(P, p_{\text{end}})$. Likewise the label of (v, u) coincides with the label of one ℓ edges of P if $A[v] \in \text{Ex}(P, u)$. Since each of the Ex sets is of cardinality ℓ, there are at least $n - 2\ell$ values for $A[v]$ for which the labels of (p_{end}, v) and (v, u) are different from the labels of ℓ edges P. Since $A[p_{\text{end}}] \neq A[u]$, for those $n - 2\ell$ values of $A[v]$ the labels of (v, u) and (p_{end}, v) are mutually distinct as well, and we have a two-edge extension of P. So (17) follows.

Obviously, the key inequality (17) can be helpful only as long as $\ell \leq n/2$. The inequalities (13)-(18) lead us to consider a Markov chain (π_k, σ_k), where $\sigma_k \in \{0, 1\}$, such that, for $1 \leq k \leq n/2$,

$$P[(\pi', \sigma') = (\pi + 1, \sigma) | \ (\pi, \sigma)] \ = \ 1 - \pi/n,$$
$$P[(\pi', \sigma') = (\pi, \sigma + 1) | \ (\pi, \sigma)] \ = \ (\pi - 1)/n,$$
$$P[(\pi', \sigma') = (\pi, \sigma) | \ (\pi, \sigma)] \ = \ 1/n,$$

if $\sigma = 0$, and

$$\begin{aligned} P[(\pi', \sigma') = (\pi, \sigma)|\, (\pi, \sigma)] \ &= \ \pi/n, \\ P[(\pi', \sigma') = (\pi+2, \sigma-1)|\, (\pi, \sigma)] \ &= \ 1 - 2\pi/n, \\ P[(\pi', \sigma') = (\pi+1, \sigma)|\, (\pi, \sigma)] \ &= \ \pi/n, \end{aligned}$$

if $\sigma = 1$.

To define the chain completely, set $\pi_1 = 1$, and $\sigma_1 = 0$. The chain terminates once $\pi_k \geq n/2$. We want to show that ℓ_k stochastically dominates π_k, that is

$$P[\ell_k > t] \geq P[\pi_k > t], \tag{19}$$

if $(\ell_1, \sigma_1) = (\pi_1, \sigma_1)$.

To this end, let us introduce the lexicographical order $\succeq$ on the pairs (ℓ, σ):

$$(\ell, \sigma) \ \succeq (\hat{\ell}, \hat{\sigma}) \ \text{ iff } \ell > \hat{\ell} \ \text{ or } \ (\ell = \hat{\ell} \ \text{ and } \sigma \geq \hat{\sigma}).$$

It is straightforward that for every state $\mathcal{S}$, and any pair (ℓ^*, σ^*)

$$P[(\ell', \sigma') \succeq (\ell^*, \sigma^*)|\, \mathcal{S}] \ \geq \ P[(\pi', \sigma') \succeq (\ell^*, \sigma^*)|\, (\pi, \sigma)]. \tag{20}$$

if $(\pi, \sigma) = (\ell(\mathcal{S}), \sigma(\mathcal{S}))$. Using this inequality and induction, one can show easily that for each k and all (ℓ_j^*, σ_j^*), $j \leq k$,

$$P[(\ell_j, \sigma_j) \succeq (\ell_j^*, \sigma_j^*), \ \forall j \leq k] \ \geq \ P[(\pi_j, \sigma_j) \succeq (\ell_j^*, \sigma_j^*), \ \forall j \leq k].$$

Setting $\sigma_j^* = 0$, we get

$$P[\ell_j \geq \ell_j^*, \ \forall j \leq k] \ \geq \ P[\pi_j \geq \ell_j^*, \ \forall j \leq k],$$

and (19) follows.

(II) Let $K = \min\{k : \ \ell_k \geq n/2\}$, and $\widetilde{K} = \min\{k : \ \pi_k \geq n/2\}$. Since ℓ_k, π_k never decrease, by (19),

$$P(K \geq k) \ \leq \ P(\widetilde{K} \geq k), \qquad \forall k.$$

To study the limiting behavior of $\widetilde{K}$, introduce $\widetilde{K}_\ell = \min\{k : \ \pi_k = n/2\}$, for $\pi_1 = \ell$, $\sigma_1 = 0$ and $\widetilde{K}_\ell^* = \min\{k : \ \pi_k = n/2\}$, for $\pi_1 = \ell$, $\sigma_1 = 1$, i.e. $\widetilde{K} = \widetilde{K}_1$ and $\widetilde{K}_{n/2} = \widetilde{K}_{n/2+1} = \widetilde{K}_{n/2}^* = \widetilde{K}_{n/2+1}^* = 0$. (For simplicity we assume that n is even.)

Introduce the Laplace transforms

$$F_\ell(u) = \mathbb{E}\big(e^{u\widetilde{K}_\ell/n}\big), \qquad F_\ell^*(u) = \mathbb{E}\big(e^{u\widetilde{K}_\ell^*/n}\big), \quad u > 0.$$

Using the Markov property of (π_k, σ_k), for $\ell < n/2$,

$$F_\ell = e^{u/n} \left[\left(1 - \frac{\ell}{n}\right) F_{\ell+1} + \left(\frac{\ell}{n} - \frac{1}{n}\right) F_\ell^* + \frac{1}{n} F_\ell \right], \tag{21}$$

$$F_\ell^* = e^{u/n} \left[\frac{\ell}{n} F_\ell^* + \left(1 - 2\frac{\ell}{n}\right) F_{\ell+2} + \frac{\ell}{n} F_{\ell+1}^* \right]. \tag{22}$$

We want to show the existence of a smooth function $h(x)$, such that $F_\ell(u), F_\ell^*(u) \sim e^{uh(\ell/n)}$. This would imply that $\widetilde{K}_\ell/n$, $\widetilde{K}_\ell^*/n$ converge, in probability, to $h(\ell/n)$. To this end, introduce also two smooth functions $a(x)$, $b(x)$ and define

$$f_\ell = e^{u(h+a/n)}, \quad f_\ell^* = e^{u(h+b/n)}, \quad h = h(\ell/n), \quad a = a(\ell/n), \quad b = b(\ell/n).$$

Our task is to determine h, a, b so that f_ℓ, f_ℓ^* *almost* satisfy the equations (21)–(22) for F_ℓ and F_ℓ^*. Plugging the expressions for f_ℓ, f_ℓ^* into these equations, and using the smoothness of $h(x)$, $a(x)$ and $b(x)$, we compute

$$\begin{aligned}
&f_\ell - e^{u/n}\left[\left(1-\frac{\ell}{n}\right) f_{\ell+1} + \left(\frac{\ell}{n}-\frac{1}{n}\right) f_\ell^* + \frac{1}{n} f_\ell\right] \\
=&f_\ell\left\{1 - e^{u/n}\left[\left(1-\frac{\ell}{n}\right) e^{uh'/n+O(n^{-2})} + \left(\frac{\ell}{n}-\frac{1}{n}\right) e^{u(b-a)/n} + \frac{1}{n}\right]\right\} \\
=&f_\ell\left\{1 - e^{u/n}\left[1 + \left(1-\frac{\ell}{n}\right)\frac{uh'}{n} + \frac{\ell}{n}\frac{u(b-a)}{n} + O\left(n^{-2}\right)\right]\right\} \\
=&-f_\ell\frac{u}{n}\left[1 + \left(1-\frac{\ell}{n}\right) h' + \frac{\ell}{n} c + O(n^{-1})\right],
\end{aligned} \tag{23}$$

where $c(x) = a(x) - b(x)$, and the bounded coefficient implicit in $O(n^{-1})$ depends on u and $\max|h''(x)|$, $\max|a'(x)|$. Likewise

$$\begin{aligned}
&f_\ell^* - e^{u/n}\left[\frac{\ell}{n} f_\ell + \left(1-2\frac{\ell}{n}\right) f_{\ell+2} + \frac{\ell}{n} f_\ell^*\right] \\
=&-f_\ell^*\frac{u}{n}\left[1 + \left(2-3\frac{\ell}{n}\right) h' + \left(1-2\frac{\ell}{n}\right) c + O(n^{-1})\right].
\end{aligned} \tag{24}$$

Interestingly, the square brackets expressions in the bottom lines of (23)–(24) depend on a and b only through the difference $c = a - b$. Let us choose $c(x)$ such that

$$\frac{1-xc(x)}{1-x} = \frac{1+(1-2x)c(x)}{2-3x} \implies c(x) = \frac{1-2x}{1-x-x^2}.$$

Then (23)–(24) are very nearly satisfied if

$$h'(x) + \frac{1-xc(x)}{1-x} = 0 \implies h'(x) + \frac{1-x}{1-x-x^2} = 0.$$

Since we want $1 = \mathbb{E}[e^{u\widetilde{K}_{n/2}/n}] \sim e^{uh(1/2)}$, we impose the condition $h(1/2) = 0$. The solution is

$$h(x) = -\frac{1}{2}\ln[4(1-x-x^2)] + \frac{3}{2\sqrt{5}}\ln\frac{(\sqrt{5}-2x-1)(\sqrt{5}+2)}{(\sqrt{5}+2x+1)(\sqrt{5}-2)}.$$

In particular,

$$h(0) = -\ln 2 + \frac{3}{2\sqrt{5}}\ln\frac{(\sqrt{5}-1)(\sqrt{5}+2)}{(\sqrt{5}+1)(\sqrt{5}-2)}.$$

Pick $\varepsilon \in (0,1)$, and set $h_\pm(x) = (1\pm\varepsilon)h(x)$, so that $h_\pm(1/2) = 0$ again. Consider the $+$-case. Let $a(x) = c(x) + M$, $b(x) = M$, $M > 0$ to be specified shortly. Then $a(x) - b(x) = c(x)$, and since $h'(x) < 0$ for $x \in [0, 0.6]$, say, the square brackets expressions on the right hand side of (23) and (24), times -1, are positive and bounded away from zero. So the corresponding f_ℓ, f_ℓ^* satisfy the recurrence *inequalities* obtained from (21)–(22) by replacing $=$ with $\le$. In addition, for $\ell \in \{n/2, n/2+1\}$,

$$f_\ell = \exp\left[u\left(h_+(\ell/n) + n^{-1}(c(\ell/n) + M)\right)\right] = \exp(un^{-1}(M - a)),$$

where $a = \max_{x\le 0.6}\bigl(2|h'(x)| + |c'(x)|\bigr)$, as $h(1/2) = c(1/2) = 0$. So, choosing $M \ge a$, we ensure that $f_\ell \ge 1$. Likewise $f_\ell^* > 1$ for those ℓ. Therefore

$$f_\ell \ge F_\ell, \quad f_\ell^* \ge F_\ell^*, \quad \ell \in \{n/2, n/2+1\}.$$

Using this as the basis of the backward induction, and the recurrence equations (inequalities respectively) for F_ℓ, F_ℓ^* (for f_ℓ, f_ℓ^* respectively) for the inductive step, we obtain that, for all $\ell \le n/2$,

$$F_\ell(u) \le \exp\left[u\bigl(h_+(\ell/n) + n^{-1}(c(\ell/n) + M)\bigr)\right], \quad F_\ell^*(u) \le \exp\left[u\bigl(h_+(\ell/n) + n^{-1}M\bigr)\right].$$

In exactly the same way we obtain the lower bounds with h_- in place of h_+, and $c(x) - M$, $-M$ in place of $c(x) + M$, M. In particular, using these bounds for $\ell = 1$, we see that $n^{-1}\widetilde{K}_1 = n^{-1}\widetilde{K}$ converges in probability, and in terms of Laplace transform, to $\alpha = h(0)$. Therefore, using the stochastic dominance of $\widetilde{K}$ over K, we obtain that for each $\delta > 0$, $K \le (1+\delta)n\alpha$ with probability approaching 1 as $n \to \infty$. Roughly, w.h.p. the number of steps (vertices) it takes for HPath to build a path of length $n/2$ is αn, at most. This proves Lemma 5.

Once such a path is determined, we switch to the algorithm in which the fresh vertices are tested in the fixed direction, clockwise or counterclockwise. For this phase the number of additional edges that will be added to the path equals, in distribution, to the total number of boxes labeled $n/2+1, \dots, n$ which are occupied by at least one ball in the uniformly random allocation of $n - K$ balls among the boxes $1, 2, \dots, n/2, n/2+1, \dots, n$. The conditional expected number of such boxes is

$$\frac{n}{2}\left[1 - \left(1 - \frac{1}{n}\right)^{n-K}\right] = \frac{n}{2}[1 - \exp(-1 + K/n + O(n^{-1})].$$

In probability,

$$\liminf(1 - e^{-1+K/n+O(n^{-1}}) \ge 1 - e^{-1+\alpha},$$

whence the expected length of the terminal path scaled by n is, in the limit, $1 - e^{-1+\alpha}$. The theorem is proved completely.

Notes. (1) The combination of Laplace transforms and the approximation technique via the differential equations had been used by second author in [P1] (an urn model), [P2] (spreading rumor process), and [P3] (random graph process).

(2) There is a natural extension of HPath algorithm in which more than one vertex can be kept alive. Numerical computations for the corresponding Markov chain indicate that the expected length of the terminal path is at least

$$\left(1 - \frac{e^{0.591-1}}{2}\right) n \approx 0.6678n,$$

a surprisingly small improvement. The original algorithm and this last modification work better in practice; to establish this fact rigorously one would need to find a better alternative to the lower bound $1 - 2\ell/n$ for the key transition probability.

(3) For n prime there is a modified algorithm that on average outperforms GPath1. In this algorithm after the k-th step we have a path $P_k = \{p_{i_0} \to p_{i_1} \to \ldots \to p_{i_\ell}\}$, $\ell = \ell(k)$, a set $D_k \supseteq P_k$ of dead vertices, and an interval $[u_k, v_k] = F_k = D_k^c$ of fresh vertices. If either $p_{i_\ell} \to u_k \to v_k$ or $p_{i_\ell} \to v_k \to u_k$ can be added to P_k, we do so, thereby getting P_{k+1} of length $\ell(k+1) = \ell(k) + 2$. If a two-edge extension is not possible, we go for one-edge extension, $p_{i_\ell} \to u_k$ or $p_{i_\ell} \to v_k$, if any is feasible, obtaining P_{k+1} of length $\ell(k+1) = \ell(k)+1$. Otherwise $P_{k+1} = P_k$. Whatever the outcome is, $D_{k+1} = D_k \cup \{u_k, v_k\}$. It is clear that $P_{k+1} = P_k$ (i.e. $\ell(k+1) = \ell(k)$) with the (conditional) probability $(\ell/n)^2$. Primality of n can be used to show that $\ell(k+1) = \ell(k) + 2$ with probability

$$\left(1 - \frac{\ell}{n}\right)\left(1 - \frac{\ell+1}{n}\right) + \frac{\lfloor (n-\ell-1)^2/4 \rfloor}{n^2}, \tag{25}$$

at least. Then $\ell(k+1) = \ell(k) + 1$ with probability

$$1 - \left(\frac{\ell}{n}\right)^2 - \left(1 - \frac{\ell}{n}\right) \cdot \left(1 - \frac{\ell+1}{n}\right) - \frac{\lfloor (n-\ell-1)^2/4 \rfloor}{n^2}$$

at most. With these bounds at hand, we construct the corresponding Markov chain $\{\ell(k)\}$ which is dominated by the second phase of GPath1. It turns out the expected terminal path length is asymptotic to $0.672n$, larger by $0.007n$ than the bound in Theorem 4. Here is the explanation for (25). First of all, $p_{i_\ell} \to u_k \to v_k$ is added to the current path with probability $(1-\ell/n)(1-(\ell+1)/n)$, regardless of whether n is prime. It remains to show that, for n prime, $p_{i_\ell} \to u_k$ cannot be added to the path, but $p_{i_\ell} \to v_k \to u_k$ can with probability $\lfloor (n-\ell-1)^2/4 \rfloor / n^2$, at least, if $\ell \geq n/2$. The latter bound follows from a theorem, due to Pollard [14]:

Theorem 6 *Let p be a prime number, and let A and B nonempty subsets of $\mathbb{Z}/p\mathbb{Z}$. Let*

$$r = |B| \leq |A| = s.$$

For $t = 1, \ldots, r$, let N_t denote the number of congruence classes in $\mathbb{Z}/p\mathbb{Z}$ that have at least t representations in the form $a + b$, where $a \in A$ and $b \in B$. Then

$$N_1 + N_2 + \ldots + N_t \geq \min\{tp, t(r+s-t)\}. \tag{26}$$

Let

$$C := \{(A[v_k], A[u_k]) : \quad p_{i_\ell} \not\to u_k, \quad p_{i_\ell} \to v_k \to u_k\}.$$

Then, clearly

$$P(p_{i_\ell} \not\to u_k, \quad p_{i_\ell} \to v_k \to u_k) \;=\; |C|/n^2. \tag{27}$$

Let

$$A := \{A[v_k] : \; p_{i_\ell} \to v_k\} \qquad \text{and} \qquad B := \{A[u_k] : \; p_{i_\ell} \not\to u_k, \quad A[u_k] \neq A[p_{i_\ell}]\}.$$

Denote $\mathcal{L}(P)$ the set of labels appearing on the edges of P. Note that $|A| = n - \ell \leq \ell$, $\ell - 1 \leq |B| \leq \ell$ and $|\mathcal{L}(P)| = \ell$. A pair of labels $(a, b) \in A \times B$ is in C, iff $a + b \pmod{n}$ is not in $\mathcal{L}(P)$. Indeed, $a \in A$ implies $p_{i_\ell} \to v_k$, $b \in B$ implies $p_{i_\ell} \not\to u_k$, and $A[u_k] \neq A[p_{i_\ell}]$ implies that the label of $p_{i_\ell} \to v_k$ is different from $a + b \pmod{n}$ (which is the label of $v_k \to u_k$).

Setting $t = \lfloor (n - \ell - 1)/2 \rfloor$, Theorem 6 can be applied for the sets A and B, as $t \leq \min\{|A|, |B|\}$, therefore by (26),

$$N_1 + N_2 + \ldots + N_t \geq \min\{tn, t(|A| + |B| - t)\} \geq t(n - 1 - t). \tag{28}$$

The left hand side of (28) counts the sums $a + b \pmod{n}$ of the pairs of $(a, b) \in A \times B$, with the restriction that a particular sum is counted at most t times. Recall that a pair of $(a, b) \in A \times B$ is in C, if $a + b \pmod{n} \notin \mathcal{L}(P)$. A particular $c \in \mathcal{L}(P)$ can occur as a sum at most t times, hence the number of sums which are not in $\mathcal{L}(P)$ is at least $N_1 + N_2 + \ldots + N_t - t \cdot |\mathcal{L}(P)|$ which is by (28) at least $t(n - \ell - 1 - t)$. Using (27), and substituting the value of T we obtain

$$\begin{aligned} P(p_{i_\ell} \not\to u_k, \quad p_{i_\ell} \to v_k \to u_k) \;&=\; |C|/n^2 \geq t(n - \ell - 1 - t)/n^2 \\ &=\; \left\lfloor \frac{n - \ell - 1}{2} \right\rfloor \cdot \left\lceil \frac{n - \ell - 1}{2} \right\rceil \cdot \frac{1}{n^2} \;=\; \frac{\lfloor (n - \ell - 1)^2/4 \rfloor}{n^2}. \end{aligned} \tag{29}$$

If we replace GPath1 with this algorithm for the second phase, when the length of the path exceeded $n/2$, then we could a little bit improve the average performance to give a path of length $0.672n$. As the analysis works only for prime n, and the improvement is marginal, we omit the details.

8 Upper bounds for harmonic paths

The paths produced by our algorithms contain a sizeable fraction of all vertices, about $2/3$ for GPath1, for instance. That this fraction is below 1, as opposed to the matching algorithm, makes it natural to ask how likely is it that the longest (non-crossing, harmonious) path has length asymptotic to n? Our next, and last, result shows that chances of this happening are exponentially small.

Proposition 7 *The probability that the length of the longest path is less than* $0.9604n$ *is* $1 - O(q^n)$*, for some* $0 < q < 1$.

Proof. Let us first compute $f(n,m)$, the total number of non-crossing paths with m vertices, $m \le n$. The number of ways to select m vertices is $\binom{n}{m}$. Picking a starting vertex (in m ways), we have two choices, left or right neighbor, for the second vertex, and recursively two choices for the j-th vertex, given the first $j-1$ vertices, $2 \le j \le m$. Therefore there are $m2^{m-2}$ *directed* paths, whence $m2^{m-3}$ undirected paths on given m vertices. Therefore

$$f(n,m) = \binom{n}{m} m2^{m-3}.$$

The probability that such a path is harmonious is

$$\frac{n \cdot n \cdot (n-1) \cdot \ldots \cdot (n-m+2)}{n^m} = \frac{(n)_{m-1}}{n^{m-1}},$$

so $\mathbb{E}(n,m)$, the expected number of non-crossing, harmonious paths with m vertices is given by

$$\mathbb{E}(n,m) = m2^{m-3}\binom{n}{m}\frac{(n)_{m-1}}{n^{m-1}}.$$

Notice that

$$\frac{\mathbb{E}(n,m+1)}{\mathbb{E}(n,m)} = \frac{2(n-m)(n-m+1)}{mn} < \frac{1}{2},$$

if $m \ge 2n/3$, say. Therefore, picking $k \ge 2n/3$,

$$\sum_{m \ge k} \mathbb{E}(n,m) \le 2\mathbb{E}(n,k).$$

Furthermore,

$$\mathbb{E}(n,k) \;=\; \frac{2^{k-3}kn}{n-k+1}\binom{n}{k}^2 \frac{k!}{n^k} \;\le\; 2^k n^3 \left[\frac{n^n}{k^k(n-k)^{n-k}}\right]^2 \left(\frac{k}{ne}\right)^k \;=\; n^3 e^{nJ(k/n)}, \tag{30}$$

where

$$J(x) = -x \cdot (1 - \ln 2) + (1-x) \cdot \ln(1-x)^{-1}.$$

Now $J(1) = -1 + \ln 2 < 0$, and $J(x)$ is decreasing on $[1/2, 1)$. The computation shows that $J(x_0) = 0$ for $x_0 = 0.96037\ldots$. Therefore

$$\mathbb{E}(n, [0.9064n]) \le n^3 e^{nJ(0.9604)} \le q^n,$$

for some $0 < q < 1$, which completes the proof. ∎

Notes. (1) Let us see what a similar computation delivers for the perfect (non-crossing, harmonious) matchings, in the case of n even of course. It is well-known (Stanley [15], Exer. 6.19 (**n**)) that the total number of non-crossing matchings is the Catalan number

$$C(n/2) = \frac{1}{n/2+1}\binom{n}{n/2} \sim c\frac{2^n}{n^{3/2}}.$$

Any such matching is harmonious with probability

$$\frac{(n)_{n/2}}{n^{n/2}} \sim c_1 \left(\frac{2}{e}\right)^{n/2},$$

and so the expected number of such matchings is asymptotic to $c_2 n^{-3/2}(8/e)^{n/2}$, thus approaching infinity exponentially fast. Whether the likely number of perfect matchings is also (exponentially) large is an interesting open problem.

(2) In fact, the Catalan number $C(n-1)$ equals the number of some particular (“alternating”) non-crossing trees with n vertices on the circle, see [15], (Exer. 6.19 ($\mathbf{p},\mathbf{q}$)), for the exact formulation and the references. So there are at least $cn^{-3/2}4^n$ non-crossing trees. And, using the depth-first traversing of any such tree, we see that it is harmonious with probability

$$\frac{(n)_{n-1}}{n^{n-1}}.$$

Therefore the expected number of harmonious, non-crossing trees is $c(4/e)^n$, at least. Could it be that w.h.p. there are exponentially many such trees?

9 Appendix

Proposition 8 *Define the random variable $X_{p,t}$ as the time it takes for the $(p, 1-p)$–random walk to reach the zero state from the state t. Then, for all $r > 0$,*

$$Pr(X_{p,t} \geq r) \leq \frac{1}{(2p)^t(4p(1-p))^{-r/2}}.$$

Proof. It is well known that the generating function for $X_{p,1}$ is

$$\mathrm{E}[z^{X_{p,1}}] \ = \ \frac{1-\sqrt{1-4p(1-p)z^2}}{2pz}, \quad |z| \ \leq \ (4p(1-p))^{-1/2}. \tag{31}$$

By the Markov property,

$$X_{p,t}\overset{\mathcal{D}}{\rightarrow} \equiv \sum_{j=1}^{t} X_{p,1}^{(j)}$$

where $X_{p,1}^{(j)}$ are independent copies of $X_{p,1}$. Therefore

$$\mathrm{E}(z^{X_{p,t}}) \ = \ \mathrm{E}^t(z^{X_{p,1}}),$$

and, for all $r > 0$,

$$\Pr(X_{p,t} \geq r) \leq \frac{\mathrm{E}^t(z^{X_{p,1}})}{z^r}, \quad \forall z \in [1, (4p(1-p))^{-1/2}].$$

Setting $z = (4p(1-p))^{-1/2}$, we obtain then

$$\Pr(X_{p,t} \geq r) \ \leq \ \frac{1}{(2p)^t((4p(1-p))^{-1/2})^{t+r}} \ \leq \ \frac{1}{(2p)^t((4p(1-p))^{-1/2})^r},$$

as claimed. ∎

Proposition 9 *For each even $t \geq 0$, let $R_p(t)$ denote the probability of being at 0 at the time step t in the (p,q)–random walk on $\{0,1,2,\ldots\}$ with the repellent 0 state. Then*

$$\begin{aligned} R_{1/2}(t) &\sim c_1 t^{-1/2}, \text{ and} \\ R_p(t) &< 3(1-2p) + (\pi t)^{-1/2}, \quad \text{if } p < 1/2. \end{aligned}$$

Let the random variable Y_p be the time it takes for the (p,q)–random walk to return to the 0 state. As above, let $X_{p,1}$ be the time it takes for the walk to reach the zero state from the state 1. Then

$$\mathrm{E}[z^{Y_p}] = z\mathrm{E}[z^{X_{p,1}}] = \frac{1-\sqrt{1-4pqz^2}}{2p}, \tag{32}$$

using (31).

Now, since

$$\sum_{r\geq 0} z^r R_p(r) = \frac{1}{1-\mathrm{E}[z^{Y_p}]},$$

we have

$$R_p(t) = [z^t]\frac{1}{1-\mathrm{E}[z^{Y_p}]}. \tag{33}$$

Then

$$R_{1/2}(t) = [z^t]\frac{1}{\sqrt{1-z^2}},$$

so that

$$R_{1/2}(t) = (-1)^{t/2}\binom{-1/2}{t/2} = 4^{-t/2}\binom{t}{t/2} \sim c_1 t^{-1/2},$$

Let us now move on to the estimation of $R_p(t)$ for $p < 1/2$. Using (32) and (33), an elementary manipulation yields

$$R_p(t) = \frac{1-2p}{q} + [z^t]\frac{2p}{\sqrt{1-4pqz^2}+(1-2p)}. \tag{34}$$

In order to estimate the second summand in the right hand side of (34), define

$$g(z) := \frac{2p}{\sqrt{1-4pqz}+(1-2p)}.$$

Now $g(z)$ is analytic in the complex plane with a cut $\{z = u+iv : v = 0,\ u \in [u_0,\infty)\}$, $u_0 = (4pq)^{-1}$, where

$$\sqrt{1-u/u_0} = e^{\pm i\pi/2}\sqrt{u/u_0 - 1},$$

with $-$ and $+$ corresponding, respectively, to the upper shore and to the lower shore of the cut. Let $\mathcal{C}$ be a (positively oriented) closed contour formed by $\mathcal{C}_1(R) = \{z = Re^{i\theta} : \theta \in$

$(0, 2\pi)\}$, and $\mathcal{C}_2(R) = \{z = u : \; u \in [R, u_0]\}$, $\mathcal{C}_3(R) = \{z = u : \; u \in [u_0, R]\}$, representing the directed lower and upper shores of the cut. Then, by Cauchy's formula,

$$[z^t]g(z) \;=\; \frac{1}{2\pi i} \oint\limits_{\mathcal{C}(R)} \frac{g(z)!}{z^{t+1}}\, dz \;=\; \frac{1}{2\pi i} \oint\limits_{\mathcal{C}_2(R)\cup\mathcal{C}_3(R)} \frac{g(z)}{z^{t+1}}\, dz + \frac{1}{2\pi i} \oint\limits_{\mathcal{C}_1(R)} \frac{g(z)!}{z^{t+1}}\, dz.$$

Let $R \to \infty$. Then the last integral tends to zero, and the first integral converges to that over $\mathcal{C}_2(\infty) \cup \mathcal{C}_3(\infty)$. Thus

$$\begin{aligned}
[z^t]g(z) =& \frac{1}{2\pi i} \int_{u_0}^{\infty} \frac{du}{u^{t+1}(e^{-i\pi/2}\sqrt{u/u_0 - 1} + 1 - 2p)} \\
&+ \frac{1}{2\pi i} \int_{\infty}^{u_0} \frac{du}{u^{t+1}(e^{i\pi/2}\sqrt{u/u_0 - 1} + 1 - 2p)} \\
=& \frac{1}{\pi} \int_{u_0}^{\infty} \frac{(u/u_0 - 1)^{1/2}}{u^{t+1}((u/u_0 - 1) + (1-2p)^2)}\, du \\
=& \frac{1}{\pi u_0^t} \int_{1}^{\infty} \frac{(y-1)^{1/2}}{y^{t+1}((y-1) + (1-2p)^2)}\, dy \\
\leq& \int_{1}^{\infty} \frac{(y-1)^{1/2}}{y^{t+1}((y-1) + (1-2p)^2)}\, dy \quad (\text{since } u_0 \geq 1) \\
=& \int_{1}^{1+(1-2p)^2} \frac{(y-1)^{1/2}}{y^{t+1}((y-1) + (1-2p)^2)}\, dy + \int_{1+(1-2p)^2}^{\infty} \frac{(y-1)^{1/2}}{y^{t+1}((y-1) + (1-2p)^2)}\, dy.
\end{aligned} \tag{35}$$

Now

$$\int_{1}^{1+(1-2p)^2} \frac{(y-1)^{1/2}}{y^{t+1}((y-1) + (1-2p)^2)}\, dy \leq (1-p)^{-2} \int_{1}^{1+(1-2p)^2} (y-1)^{1/2}\, dy \leq 1 - 2p, \tag{36}$$

and

$$\begin{aligned}
\int_{1+(1-2p)^2}^{\infty} \frac{(y-1)^{1/2}}{y^{t+1}((y-1) + (1-2p)^2)}\, dy &\leq \int_{1}^{\infty} \frac{dy}{y^{t+1}(y-1)^{1/2}} \quad (y = e^u) \\
&= \int_{0}^{\infty} e^{-ut}(e^u - 1)^{-1/2}\, du \leq \int_{0}^{\infty} e^{-ut}u^{-1/2}\, du = (\pi t)^{-1/2}.
\end{aligned} \tag{37}$$

Using (35), (36), and (37), we obtain

$$[z^t]g(z) \leq 1 - 2p + (\pi t)^{-1/2}$$

Therefore, by (34) we have

$$R_p(t) \leq \frac{1-2p}{q} + 1 - 2p + (\pi t)^{-1/2} < 3(1-2p) + (\pi t)^{-1/2},$$

since $q > 1/2$. ∎

Acknowledgements: We would like to thank Gyula Károlyi for drawing our attention to Pollard's result.

References

[1] M. Abellanas, J. García, G. Hernández, M. Noy, and P. Ramos, Bipartite embeddings of trees in the plane, in: *Graph Drawing* (S. North, ed.), *Lecture Notes in Computer Science* **1190**, Springer–Verlag, Berlin, 1997, 1–10. Also in: *Discrete Applied Math.* **93** *(1999) 141–148.*

[2] N. Alon, S. Rajagopalan, S. Suri, Long non-crossing configurations in the plane. *Fund. Inform.* **22** (1995) 385–394.

[3] G. Araujo, J. Balogh, R. Fabila, G. Salazar, J. Urrutia, Harmonic subgraphs in labeled geometric graphs, submitted.

[4] J.A. Gallian, A dynamic survey of graph labeling. *Electron. J. Combin.* **5** (1998).

[5] R. L. Graham and N. J. A. Sloane, On additive bases and harmonious graphs, *SIAM J. Alg. Discrete Meth.,* **1** (1980).

[6] C. Hernando, F. Hurtado, M. Noy, Graphs of non-crossing perfect matchings. *Graphs Combin.* **18** (2002) 517–532.

[7] A. Kaneko and M. Kano, Discrete Geometry on red and blue points in the plane — a survey, *Discrete and Computational Geometry* (B. Aronov et al., eds.), Springer–Verlag, Berlin, 2004, 551–570.

[8] A. Kaneko, M. Kano, and K. Yoshimoto, Alternating Hamiltonian cycles with minimum number of crossings in the plane, *Internat. J. Comput. Geom. Appl.* **10** (2000), 73–78.

[9] J. Kÿncl, G. Tóth, and J. Pach, Long alternating paths in bicolored point sets. Preprint (2004).

[10] T. Łuczak, B. Pittel, and J. C. Wierman, The structure of a random graph at the point of the phase transition, *Trans. Amer. Math. Soc.* **341** (1994), 721–748.

[11] B. Pittel, An urn model for cannibal behavior, *J. Appl. Prob.* **24** (1987) 522–526.

[12] B. Pittel, On a Daley-Kendall model of random rumors, *J. Appl. Prob.* **27** (1990) 14–27.

[13] B. Pittel, On tree census and the giant component in sparse random graphs, *Random Structures and Algorithms* **1** (1990) 311–342.

[14] J. M. Pollard, A generalization of a theorem of Cauchy and Davenport, *J. London Math. Soc.* **8** (1974) 460–462.

[15] R. P. Stanley, Enumerative Combinatorics, 2, *Cambridge Studies in Advanced Mathematics 62* (1999).

[16] S. Tokunaga, Intersection number of two connected geometric graphs, *Inform. Process. Lett.* **59** (1996) 331–333.