Permutations resilient to deletions

Noga Alon, Steve Butler, Ron Graham and Utkrisht C. Rajkumar

Abstract. Let $M = (s_1, s_2, \ldots, s_n)$ be a sequence of distinct symbols and σ a permutation of $\{1, 2, \ldots, n\}$. Denote by $\sigma(M)$ the permuted sequence $(s_{\sigma(1)}, s_{\sigma(2)}, \ldots, s_{\sigma(n)})$. For a given positive integer d, we will say that σ is *d-resilient* if no matter how d entries of M are removed from M to form M' and d entries of $\sigma(M)$ are removed from $\sigma(M)$ to form $\sigma(M)'$ (with no symbol being removed from both sequences), it is always possible to reconstruct the original sequence M from M' and $\sigma(M)'$. Necessary and sufficient conditions for a permutation to be d-resilient are established in terms of whether certain auxiliary graphs are acyclic. We show that for d-resilient permutations for $[n]$ to exist, n must have size at least exponential in d, and we give an algorithm to construct such permutations in this case. We show that for each d and all sufficiently large n, the fraction of all permutations on n elements which are d-resilient is bounded away from 0.

1. Introduction

Suppose we are trying to send a message M consisting of n distinct symbols over some *deletion* channel. This channel has the property that for some fixed d, at most d symbols might possibly be deleted, and that in the resulting message, M', the remaining symbols are concatenated so that the positions of the deleted symbols are not given. Of course, just seeing M', we have no way of knowing what (or where) any deleted symbol was in M. To deal with this problem, we take a typical coding theory approach and transmit additional information. In particular, we will transmit a permutation $\sigma(M)$ of M over the channel, and see if M' and $\sigma(M)'$ are enough to reconstruct the original message M. An obvious necessary condition is the channel doesn't delete the same symbol in M and $\sigma(M)$.

.

(Alon) Research supported in part by a BSF grant, an ISF grant and a GIF grant.
(Butler) Partially supported by a grant from the Simons Foundation (#427264).

We will say that the permutation σ on $[n] := \{1, 2, \ldots, n\}$ is *d-resilient* if it is always possible to reconstruct the original message M from the two resulting messages M' and $\sigma(M)'$ (provided that no symbol is missing from both resulting messages).

For example, consider the permutation $\sigma = \left(\begin{smallmatrix}1&2&3&4\\4&1&3&2\end{smallmatrix}\right)$. If the original message is $M = ABCD$ then the permuted message is $\sigma(M) = DACB$. If the channel deletes the symbol C in M and B in $\sigma(M)$, then the two received messages are $M' = ABD$ and $\sigma(M)' = DAC$. However, if instead, the original message were $N = ACBD$ so that $\sigma(N) = DABC$, and C is deleted in N and B is deleted in $\sigma(N)$ then we would have $N' = ABD = M'$ and $\sigma(N)' = DAC = \sigma(M)'$. Thus the permutation σ is **not** 1-resilient. However, it is easy to check that the permutation $\sigma = \left(\begin{smallmatrix}1&2&3&4\\3&1&4&2\end{smallmatrix}\right)$ is 1-resilient. More generally, it not difficult to show the following:

Fact 1.1. *A permutation σ on $[n]$ is 1-resilient if and only if consecutive elements in $[n]$ are not consecutive in σ, i.e., $|\sigma(i)-\sigma(i+1)| > 1$ for $1 \le i < n$.*

We give a necessary and sufficient condition for a permutation σ to be d-resilient expressed in terms of a family of auxiliary graphs being acyclic (see Section 2). We also give a construction of d-resilient permutations which have size n exponential in d, and show that this growth rate is best possible. Moreover we show that for every fixed d and large n a positive fraction (independent of n) of all the permutations of n elements are d-resilient (see Section 3).

Comment. Our problem was inspired in part by the oral transmission protocols for Sanskrit literature in the Vedic period. This relied on interleaving patterns of words to combat transpositions, substitutions, insertions, and deletions of words.

2. A necessary and sufficient condition to be d-resilient

We start with an example. Let $\sigma = \left(\begin{smallmatrix}1&2&3&4&5&6&7&8&9\\7&4&1&8&5&2&9&6&3\end{smallmatrix}\right)$. Suppose that we have received $ABCEGHI$ for M' and $DAHEBIF$ for $\sigma(M)'$. The symbols A, B, E, H, and I *doubly occurring*, that is they appear in both received messages. Suppose that we have been able to determine the location of the doubly occurring symbols, so that the situation is as illustrated in Figure 1. We have marked with a line how σ connects the entries involved in a deletion in the first line (M') and the second line ($\sigma(M)'$).

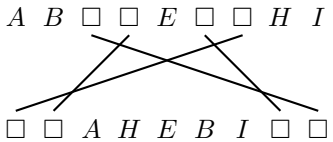

FIGURE 1. Intermediate step in working to recover M.

The entries marked as □ in Figure 1 indicate that this is a location of a symbol that occurs only once. Examining the location of these we see that there are four blocks of contiguous □'s (two in the top line and two in the bottom line). By examining M' and $\sigma(M)'$ we can conclude that each one of these blocks has one entry which was deleted and the other was transmitted. If we could uniquely determine which entries in all blocks were deleted and which were transmitted we could recover our message (we use the connections to fill any gaps). On the other hand, it might be that there is more than one possibility to which entries in the blocks were deleted and which were transmitted.

We are in the latter case in that there are two ways in which entries could have been deleted, as shown in Figure 2. Here we have oriented the edges from where a symbol was deleted (marked with a "$*$") to where it was transmitted. This allows us to determine two possible messages for M, namely $ABCDEFGHI$ and $ABFCEGDHI$.

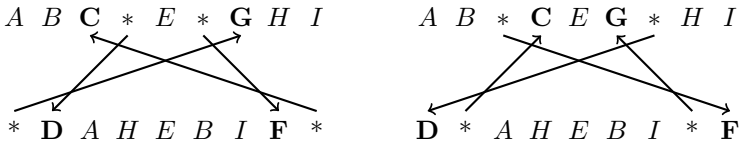

FIGURE 2. Ambiguity found in working to recover M.

We introduce an auxiliary (multi-)graph to detect such ambiguities. Given a permutation $\sigma = \left(\begin{smallmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{smallmatrix}\right)$ on $[n] = \{1, 2, \ldots, n\}$ and $D \subseteq [n]$ (representing the indices involved in deletions), let $G(\sigma, D)$ be a bipartite (multi-)graph defined in the following way. Let $t_1, t_2, \ldots, t_i$ be the sets of indices occurring in maximal contiguous blocks of elements of D in the top line of σ and let $b_1, b_2, \ldots, b_j$ be the set of indices occuring in maximal contiguous blocks of elements of D in the bottom line of σ. We now let

$$V\big(G(\sigma, D)\big) = \{t_1, t_2, \ldots, t_i, b_1, b_2, \ldots, b_j\}$$

and we add $|t_k \cap b_\ell|$ edges joining t_k and b_ℓ for all k and ℓ. Note that $G(\sigma, D)$ will have $|D|$ edges (i.e., one edge for each element in D).[1]

Since every edge in $G(\sigma, D)$ can be identified with an element of D, we can indicate whether a symbol is deleted in the top or bottom line by orienting the edge *away* from where the deletion occurs.

Returning to our example shown in Figure 1 we have $D = \{3, 4, 6, 7\}$, $t_1 = \{3, 4\}$, $t_2 = \{6, 7\}$, $b_1 = \{7, 4\}$, $b_2 = \{6, 3\}$, and $G(\sigma, D)$ is a (simple) 4-cycle. We can interpret the situation shown in Figure 2 as coming from two different cyclic orientations of the 4-cycle.

[1] This process is similar to what is done to construct random graphs with prescribed degrees; namely, we have a matching (the connections joining indices in the top and bottom lines), and then we group a cluster of endpoints together to form a vertex. Here our clusters are defined by the contiguous blocks.

Theorem 2.1. *If $|D| \leq 2d$ and $G(\sigma, D)$ has a cycle, then σ is not d-resilient.*

Proof. It suffices to show how to use the cycle to produce two distinct M for a given M' and $\sigma(M)'$. To do this carry out the following.

- Orient the edges of the cycle to produce in-degree and out-degree one at each vertex of the cycle.
- Orient the *remaining* edges so that there are at most d edges directed into the $\{t_k\}$ collectively and at most d edges directed into the $\{b_\ell\}$ collectively.

Call this orientation H_1. Let H_2 be the orientation found by starting with H_1 and reversing all the edges of the cycle. The two orientations will indicate how to delete entries in the messages M and $\sigma(M)$, i.e. we delete the entries which correspond to the tails of the directed arcs.

Let M_1 be the message $12\ldots n$, we now create a second message $M_2 \neq M_1$ so that $M_1' = M_2'$ and $\sigma(M_1)' = \sigma(M_2)'$ by doing the following.

1. Place the directed edges of the orientation of H_1 between the two lines coming from the the two-line representation of σ. For any element that has an edge oriented out replace the symbol with a $*$.
2. Replace the orientation H_1 with the orientation H_2.
3. For each block t_k and b_ℓ, move the non-$*$ entries to correspond to the vertices with an edge directed in; while preserving the relative order of the entries. The entries with edges directed out will now obtain a $*$.
4. Replace any symbol with a $*$ by using the edges of the orientation, i.e., with what it connects with in the other line.

The key step is the third step, because we have guaranteed two things to happen. First we have changed the orientation of at least two edges (from the cycle) and thus the location of at least two entries in the message has changed. Second if we delete the original representation as indicated by H_1 and the new representation as indicated by H_2 then they will produce the same pair of received messages. □

Theorem 2.2. *If for all $|D| \leq 2d$ the graph $G(\sigma, D)$ is acyclic, then σ is d-resilient.*

Proof. First we demonstrate that we can determine the location of all doubly occurring symbols in σ.

Suppose that the symbol x occurs in M' in position q. Then x must be in one of positions q, $q+1$, $\ldots$, $q+d$ in M (i.e., it could move down by at most d entries); which in turn implies that the location of x is in one of $d+1$ possible positions in $\sigma(M)$. It now suffices to show that these positions are at pairwise distance of more than d apart, since then the positions are associated with non-overlapping portions of $\sigma(M)'$ from which we can then determine the location of x. (Here distance is the difference of the indices giving the location of the entries.)

So suppose that some pair of positions are at pairwise distance d or less apart. Then there are a pair of symbols y and z so that the distance between

them in both the top and bottom lines in the two-line representation of σ is at most distance d. Now form the set D by taking y, z, and all elements between y and z in both the top and bottom lines. This has size $|D| \leq 2d$ and the vertices y and z are both in the same block on the top and bottom. Thus $G(\sigma, D)$ would have a two-cycle, which contradicts our assumption.

Since we now know the locations of all the doubly occurring symbols, we also know the locations of entries involved in deletion, i.e., there is a unique D associated with M' and $\sigma(M)'$. We also know that $G(\sigma, D)$ does not contain a cycle. We now observe that if there were two (or more) possible M, then they would have to correspond to two distinct orientations, say H_1 and H_2, of the edges of $G(\sigma, D)$, and moreover that the orientations would have the same in- and out-degrees at each vertex. (This last statement follows from noting that we know how many symbols were in these blocks from M' and $\sigma(M)'$ and also the lengths of the blocks.)

So suppose there were two possible M and let e_1 be any edge of H_2 which has a different orientation from H_1. Now going into the vertex it is oriented towards there must be some other edge that initially was oriented into the vertex in H_1 but is now oriented out. Call that edge e_2. Now we can repeat this procedure finding e_1, e_2, e_3, ..., until we eventually come to an edge which goes into a previously seen vertex by this procedure. But at such a point we have a directed cycle in H_2, and more importantly a cycle in $G(\sigma, D)$, which is impossible. So there can only be one M. Since this is true for any M' and $\sigma(M)'$ we have σ is d-resilient. □

3. Construction of d-resilient permutations

By Theorems 2.1 and 2.2 it is now easy to establish the fact from the introduction, namely that σ is 1-resilient if and only if the permutation does not map adjacent entries to adjacent entries. This first happens with $n = 4$, for example, with $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{smallmatrix}\right)$. Through an exhaustive search it was determined that the smallest possible 2-resilient permutations have length 18. One such example is:

$$\left(\begin{array}{cccccccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\ 6 & 16 & 9 & 1 & 5 & 11 & 15 & 2 & 7 & 12 & 17 & 4 & 8 & 14 & 18 & 10 & 3 & 13 \end{array}\right). \tag{3.1}$$

The following result shows that d-resilient permutations exist for every d and there is also an efficient procedure to construct them.

Theorem 3.1. *For any n and d satisfying $n > 3^{2d}$ there is a d-resilient permutation σ of $[n]$. Such a σ can be found by a polynomial time algorithm (in n).*

Call a graph H an (n, d)-double path graph if it has n vertices, its edge set is a union of two Hamiltonian paths, and its girth is at least $2d+1$. Given such a graph, number its vertices by the integers $1, 2, 3, \ldots, n$ according to the order of the first Hamiltonian path (corresponding to the top row of the two-line representation of σ). The ordering of the second Hamiltonian path

will then correspond to the bottom row of the two-line representation of σ. As an example the graph shown in Figure 3 produces the permutation in (3.1).

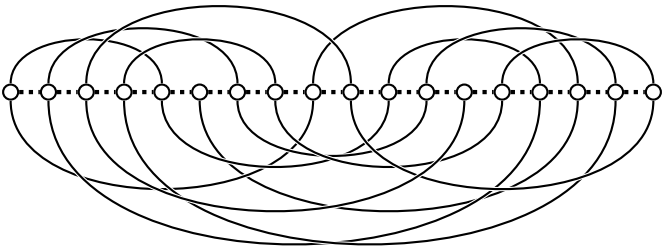

FIGURE 3. A graph corresponding to the permutation in (3.1); the dashed line being the top row and the solid line being the bottom row. Note the graph has girth 5.

Lemma 3.2. *For any (n, d)-double path graph H, the corresponding permutation is d-resilient.*

Proof. If there is a $D \subset [n]$ so that $G(\sigma, D)$ contains a cycle, then so does the induced subgraph of H on D. Since H has girth $2d + 1$ then $G(\sigma, D)$ is acyclic for all $|D| \leq 2d$ and so by Theorem 2.2 we have σ is d-resilient. □

Lemma 3.3. *If $n > 3^{2d}$ then there is an (n, d)-double path graph H.*

Proof. We apply a variant of a method of Erdős and Sachs [2]. Starting with a graph H on the set of vertices $[n]$ with the edge set being the union of the Hamiltonian path $1, 2, \ldots, n$ (in this order) and another Hamiltonian path P, we keep modifying P as long as there is a cycle of length at most $2d$ in H. We show how to perform these modifications in order to get rid of all cycles of length at most $2d$ keeping the first Hamiltonian path and maintaining the property that the second one, P, also stays a Hamiltonian path. In each modification we switch some pair of edges of P which are far from each other, that is, omit them and connect their endpoints by new edges in the unique way ensuring that the modified P will stay a Hamiltonian path. Here are the details.

As long as H contains a cycle of length at most $2d$, let C be a shortest cycle in H, and let e be an arbitrary edge of P that belongs to C (there must be such an edge, as the other Hamiltonian path contains no cycle at all). By assumption

$$n - 1 > 1 + 2 \cdot 3 + 2 \cdot 3^2 + 2 \cdot 3^3 + \cdots + 2 \cdot 3^{2d-1} = 3^{2d} - 2.$$

Starting with any edge e (corresponding to the 1 in the above sum) there are $2 \cdot 3$ edges which are incident with that edge, at most $2 \cdot 3^2$ edges which are

distance two from that edge, etc. This implies there are at most $3^{2d}-2 < n-1$ edges which are within distance $2d-1$ from e and so P contains an edge e' whose distance (in the graph H) from e is at least $2d$. We now switch at e, e' (that is, delete them and add the required edges to keep P a Hamiltonian path). This way we get rid of the cycle C. Any new cycle created this way either contains only one of the new edges, and then its length is at least $2d+1$, or contains both and then its length is at least twice the length of the shortest cycle deleted. Proceeding in this way we increase the length of the shortest cycle after finitely many steps. When the process terminates we get the required graph. Note that since the number of cycles of length t in a graph of maximum degree 4 and n vertices is smaller than $n \cdot 3^t$ the process terminates after at most

$$O\big(n(3^3 + \cdots + 3^{2d})\big) = O(n^2)$$

steps, and each step is efficient since finding a shortest cycle in a graph can be done in polynomial time. □

The assertions of Theorem 3.1 follow from the two preceding lemmas. This shows that we can find d-resilient permutations which have size n only exponential in d. We now show that this is best possible.

Theorem 3.4. *If there is a permutation σ of $[n]$ that is d-resilient, then $d \leq O(\log n)$.*

Proof. By the known results about cycles in graphs with n vertices and $2n-2$ edges (see [1]), the graph constructed from the permutation σ as in the discussion above contains a short cycle of length at most $2\log_3 n + O(1)$. This in turn implies that $G(\sigma, D)$ has a cycle of length $2\log_3 n + O(1)$ and hence we have that $d \leq 2\log_3 n + O(1)$. □

Finally we note that asymptotically a positive portion of permutations are d-resilient.

Proposition 3.5. *For any fixed d there is a positive real $\epsilon(d)$ and $n_0 = n_0(d)$ so that the probability that a random permutation σ of $[n]$ is d-resilient is at least $\epsilon(d)$.*

Proof. It is known that for every fixed integer d, a random 4-regular graph on n vertices, for large n, has girth larger than $2d$ with probability at least some $\delta(d) > 0$. By the known results about contiguity (see [3]) this random graph is the edge disjoint union of two Hamiltonian cycles with probability that tends to 1 as n tends to infinity. This implies the required assertion, by Lemma 3.2. □

Acknowledgements

Utkrisht Rajkumar thanks Young-Han Kim for support and guidance. Noga Alon and Ron Graham thank the Simons Institute for the Theory of Computing at UC Berkeley, where part of this work was done. The authors thank the referees for feedback on an earlier version of this paper.

References

[1] N. Alon, S. Hoory and N. Linial, The Moore bound for irregular graphs, Graphs and Combinatorics 18 (2002), 53-57.

[2] P. Erdős and H. Sachs, Reguläre Graphen gegebener Taillenweite mit minimaler Knotenzahl, Wiss. Z. Matin-Luther-Univ. Halle-Wittenberg Math.-Natur. Reihe 12(1963), 251–257.

[3] J. H. Kim and N. Wormald, Random matchings which induce Hamiltonian cycles, and hamiltonian decompositions of random regular graphs, J. Combinatorial Theory, Series B 81 (2001), 20–44.

Noga Alon
Sackler School of Mathematics and Blavatnik
School of Computer Science
Tel Aviv University, Tel Aviv
Israel
e-mail: nogaa@tau.ac.il

Steve Butler
Dept. of Mathematics
Iowa State University
Ames, IA 50011
USA
e-mail: butler@iastate.edu

Ron Graham
Dept. of Computer Science and Engineering
UC San Diego
La Jolla, CA 92093
USA
e-mail: graham@ucsd.edu

Utkrisht C. Rajkumar
Dept. of Computer Science and Engineering
UC San Diego
La Jolla, CA 92093
USA
e-mail: urajkuma@eng.ucsd.edu