# Sorting by Reversals

Jonas Hübotter

July 10, 2020

# Outline

## Definition 1

We define the symmetric group $\langle S_n, \circ \rangle$ as the group whose elements are all bijections over $[1, n]$ with

$$S_n = \{(0\ \pi_1\ \dots\ \pi_n\ n+1) \mid \{\pi_1, \dots, \pi_n\} = [1, n]\}$$

where $\pi_i = \pi(i)$, $\pi_0 = 0$, and $\pi_{n+1} = n+1$.

$\pi \in S_n$ is a permutation.

$id = (0\ 1\ \dots\ n\ n+1) \in S_n$ is the *identity permutation*.

Definition 2

A reversal $\rho(i,j) \in S_n$ is defined as

$$\rho(i,j) = (0\ 1\ \cdots\ i-1\ j\ j-1\ \cdots\ i+1\ i\ j+1\ \cdots\ n\ n+1)$$

for some $i, j \in [1, n]$ with $j \geq i$.

Example

Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$.
Then

$$\pi \circ \rho(2,4) = (0\ 1\ 2\ 4\ 3\ 5).$$

Definition 3 (reversal distance problem)

Given two permutations $\sigma, \tau \in S_n$ find a sequence of reversals $\rho_1, \ldots, \rho_d \in S_n$ such that

$$\sigma \circ \rho_1 \circ \cdots \circ \rho_d = \tau$$

and $d$ is minimal.

$d$ is called reversal distance between $\sigma$ and $\tau$.

Observation: The reversal distance between $\sigma$ and $\tau$ is the same as the reversal distance between $\tau^{-1} \circ \sigma$ and $id$.

Definition 4 (MIN-SBR)

Let $\pi = \tau^{-1} \circ \sigma \in S_n$.
Sorting by Reversals is the problem of finding a sequence of reversals $\rho_1, \ldots, \rho_d \in S_n$ such that

$$\pi \circ \rho_1 \circ \cdots \circ \rho_d = id$$

and $d$ is minimal.

Example

Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$.

$$\pi \circ \rho(2,4) = (0\ 1\ 2\ 4\ 3\ 5)$$
$$\pi \circ \rho(2,4) \circ \rho(3,4) = (0\ 1\ 2\ 3\ 4\ 5) = id$$

$\implies d(\pi) \leq 2$.

A different perspective: $\pi = (0\ 1 \mid 3\ 4 \mid 2 \mid 5)$

Definition 5

Let $i \sim j$ if $|i - j| = 1$.
A pair of consecutive elements $\pi_i$ and $\pi_j$ is

- an adjacency if $\pi_i \sim \pi_j$; and
- a breakpoint if $\pi_i \not\sim \pi_j$.

$b(\pi)$ denotes the number of breakpoints in $\pi \in S_n$.

Observation: $b(\pi) = 0$ iff $\pi = id$ and any reversal can at most eliminate two breakpoints.

Corollary 6 (lower bound, Kececioglu et al.)

$d(\pi) \geq \left\lceil \frac{b(\pi)}{2} \right\rceil$ *for all* $\pi \in S_n$.

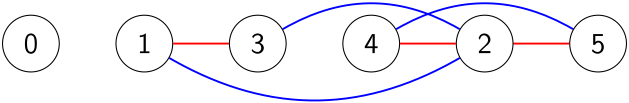
Definition 7 (breakpoint graph, Bafna et al.)

Let $G(\pi) = (V, E)$ with

- vertices $V = [0, n+1]$ representing the elements of $\pi$; and
- edges $E = R \cup B$ with
  - a red edge for every breakpoint in $\pi$; and
  - a blue edge for every missing adjacency in $\pi$.

Example

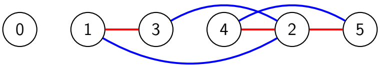
Let $\pi = (0\ 1 \mid 3\ 4 \mid 2 \mid 5) \in S_4$. Then $G(\pi)$ is




Observation: Each vertex has an equal number of incident red and blue edges.

Corollary 8 (Bafna et al.)

*$G(\pi)$ can be decomposed into edge-disjoint alternating cycles.*




Definition 9

A reversal is called *$k$-reversal* if it removes $k$ breakpoints.

A reversal *acts on* two red edges of $G(\pi)$ if those two edges represent the breakpoints that are split apart by the reversal.

An alternating cycle in $G(\pi)$ is a *$k$-cycle* if it has $k$ constituting red edges.

We call an alternating cycle $C$ in $G(\pi)$ *oriented* if there is a 1- or 2-reversal acting on two constituting red edges of $C$.

Let $c(\pi)$ denote the maximum number of alternating cycles in any alternating cycle decomposition of $G(\pi)$.

Theorem 10 (Bafna et al.)

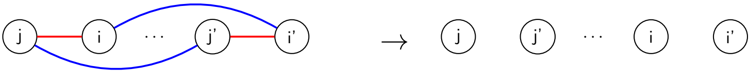
*Let $\pi, \rho \in S_n$ and $\rho$ be a reversal. Then*

$$b(\pi) - b(\pi \circ \rho) + c(\pi \circ \rho) - c(\pi) \leq 1.$$

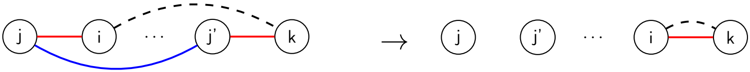
Proof.

To show: $b(\pi) - b(\pi \circ \rho) + c(\pi \circ \rho) - c(\pi) \leq 1$.
We consider each case $b(\pi) - b(\pi \circ \rho) \in [-2, 2]$ separately.

1. A 2-reversal removes at least one alternating cycle from the maximum alternating cycle decomposition.




2. A 1-reversal does not add an alternating cycle to the maximum alternating cycle decomposition.




Proof for other cases similar. □

Theorem 11 (lower bound, Bafna et al.)

*Let $\pi \in S_n$. Then*

$$d(\pi) \geq b(\pi) - c(\pi).$$

Proof.

Let $\pi_t = \pi, \pi_0 = id$ and $\rho_1, \dots, \rho_t$ a shortest sequence of reversals from $\pi_t$ to $\pi_0$. Then

$$\begin{aligned} d(\pi_i) &= d(\pi_{i-1}) + 1 \\ &\overset{(10)}{\geq} d(\pi_{i-1}) + b(\pi_i) - b(\pi_{i-1}) + c(\pi_{i-1}) - c(\pi_i) \end{aligned}$$

$$\begin{aligned} \Longleftrightarrow d(\pi_i) - (b(\pi_i) - c(\pi_i)) &\geq d(\pi_{i-1}) - (b(\pi_{i-1}) - c(\pi_{i-1})) \\ &\geq d(\pi_0) - (b(\pi_0) - c(\pi_0)) = 0 \end{aligned}$$

Setting $i = t$, proves the theorem. □

Theorem 12 (lower bound with 2-cycles, Christie)

*Let $\pi \in S_n$ and $\mathcal{C}$ be a maximum alternating cycle decomposition of $G(\pi)$. Let $c_2(\pi)$ be the minimum number of alternating 2-cycles in any such $\mathcal{C}$. Then*

$$d(\pi) \geq \frac{2}{3} b(\pi) - \frac{1}{3} c_2(\pi).$$

## $\frac{3}{2}$-approximation

By theorem 12, an algorithm that finds a sorting sequence of reversals of at most length $b(\pi) - \frac{1}{2}c_2(\pi)$ achieves an approximation bound of $\frac{3}{2}$.

We find such an algorithm in two steps:

1. given an alternating cycle decomposition $\mathcal{C}$ of $G(\pi)$ we find a sorting sequence of reversals for $\pi$; and
2. we find an alternating cycle decomposition of $G(\pi)$ maximizing the number of 2-cycles.

Lastly, we prove the approximation bound.

Definition 13 (reversal graph, Christie)

Given an alternating cycle decomposition $\mathcal{C}$ of $G(\pi)$, define $R(\mathcal{C})$ with

- an isolated blue vertex for each adjacency in $\pi$;
- $m$ vertices for each $m$-cycle in $\mathcal{C}$ each representing the reversal $\rho(u)$ acting on the two red edges connected by a blue edge;
    - a vertex is red if the represented reversal is a 1- or 2-reversal
    - a vertex is blue otherwise
- connect two vertices with an edge if their corresponding blue edges *interleave*.

Observation: The only alternating cycle decomposition of $G(id)$ is $\mathcal{C} = \emptyset$.

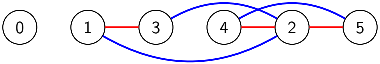
Corollary 14 (Christie)

*$R(\emptyset)$ consists of $n + 1$ isolated blue vertices.*

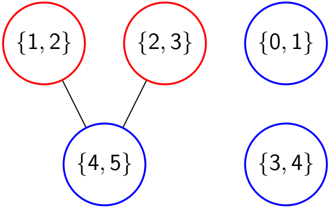
## Example

Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$.
Given the alternating cycle decomposition $\mathcal{C}$ of $G(\pi)$

$$\mathcal{C} = \{(\{1,3\},\{2,3\},\{2,4\},\{4,5\},\{2,5\},\{1,2\})\}$$




construct $R(\mathcal{C})$.




Idea: Each connected component of $R(\mathcal{C})$ can be sorted separately.

Let $u$ be a vertex of $R(\mathcal{C})$ representing reversal $\rho(u)$.

**Definition 15**

Denote by $\mathcal{C}_u$ the alternating cycle decomposition of $G(\pi \circ \rho(u))$ that is obtained from $\mathcal{C}$.

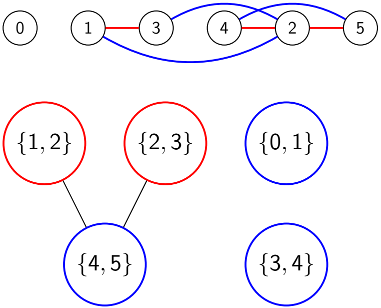
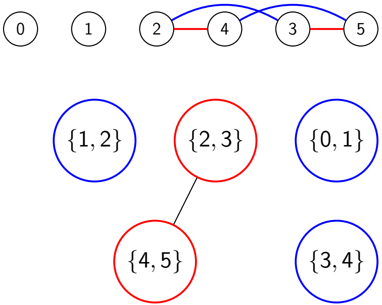
**Lemma 16 (Christie)**

*$R(\mathcal{C}_u)$ can be derived from $R(\mathcal{C})$ by making the following changes to $R(\mathcal{C})$:*

1. *flip the color of every vertex adjacent to $u$;*
2. *flip the adjacency of every pair of vertices adjacent to $u$; and*
3. *if $u$ is a red vertex, turn it into an isolated blue vertex.*

## Example

Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$ and $u = \{1, 2\}$.

| $G(\pi)$ and $R(\mathcal{C})$ | $G(\pi \circ \rho(u))$ and $R(\mathcal{C}_u)$ |
|---|---|
|  |  |

Lemma 17 (Christie)

*All vertices arising from the same alternating cycle in $\mathcal{C}$ are in the same connected component of $R(\mathcal{C})$.*

Lemma 18 (Christie)

*Vertices arising from an unoriented 2-cycle of $\mathcal{C}$ must be in a connected component of $R(\mathcal{C})$ with vertices arising from at least one more alternating cycle of $\mathcal{C}$.*

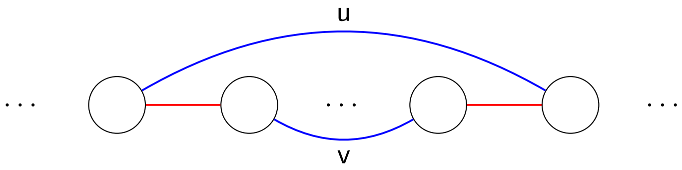



Figure 1: Unoriented 2-cycle

Definition 19

We call a connected component of $R(\mathcal{C})$ oriented if it contains a red vertex or if it consists solely of an isolated blue vertex.

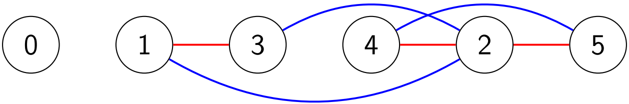
Let $A$ be a connected component of $R(\mathcal{C})$. We denote by $A_u$ the subgraph of $R(\mathcal{C}_u)$ that contains all the vertices of $A$.

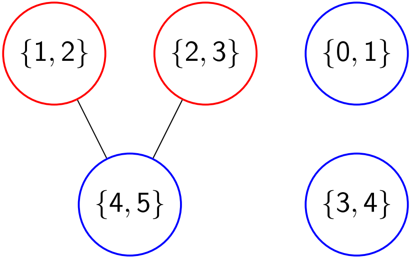
Lemma 20 (Christie)

*If a connected component $A$ of $R(\mathcal{C})$ is oriented and not an isolated blue vertex, it contains a red vertex $u$ such that $A_u$ is still oriented.*

Example (elimination sequence)

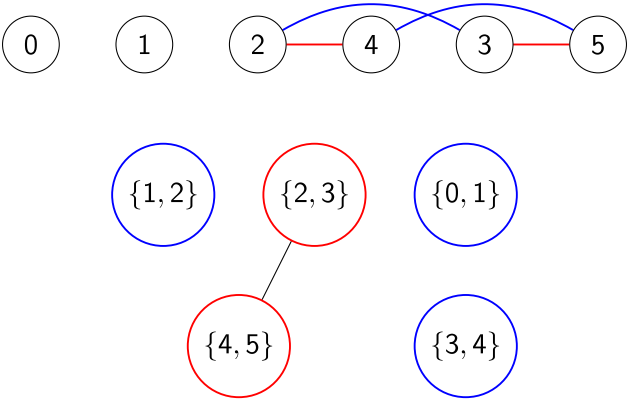
Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$



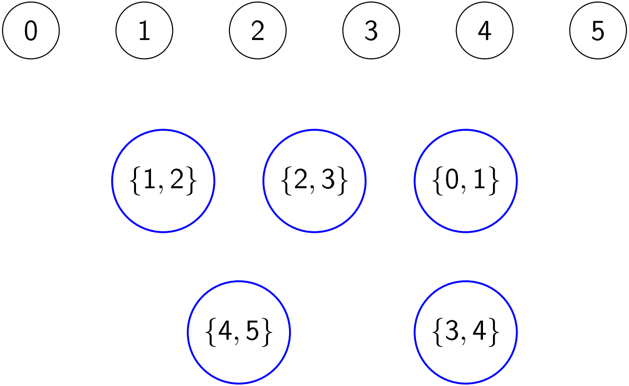
Example (elimination sequence)

Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$ and $u_1 = \{1, 2\}$



Example (elimination sequence)

Let $\pi = (0\ 1\ 3\ 4\ 2\ 5) \in S_4$ and $u_1 = \{1,2\}$, $u_2 = \{2,3\}$.



Observations:

- Every connected component $A$ arising from $k$ different alternating cycles of $G(\pi)$, eventually reduces to $k$ 2-cycles.
- There exists an elimination sequence of $A$ with $k$ 2-reversals and remaining 1-reversals.
- An unoriented connected component requires one initial 0-reversal.

**Theorem 21 (Christie)**

*Let $\pi \in S_n$ and $\mathcal{C}$ be an alternating cycle decomposition of $G(\pi)$. Then*

$$d(\pi) \leq b(\pi) - |\mathcal{C}| + u(\mathcal{C})$$

*where $u(\mathcal{C})$ is the number of unoriented components in $R(\mathcal{C})$.*

Goal: Find a cycle decomposition of $G(\pi)$ that has a large number of 2-cycles.

Idea 22

1. Construct a matching graph $F(\pi)$ where vertices represent red edges in $G(\pi)$ and vertices $u, v$ are adjacent if they form a 2-cycle in $G(\pi)$.
2. Find maximum cardinality matching $M$ of $F(\pi)$.
3. Use a ladder graph $L(M)$ with vertices representing 2-cycles in $M$ and form connected components (*ladders*) with 2-cycles sharing a blue edge in $G(\pi)$.

Definition 23

We call a 2-cycle selected if its corresponding edge of $F(\pi)$ is in $M$.

A selected 2-cycle is called independent if it is not part of a ladder. Otherwise it is called a ladder 2-cycle.

Let $z$ be the number of independent 2-cycles, and $y$ the number of ladder 2-cycles.

Theorem 24 (Christie)

*Given a maximum cardinality matching $M$ of $F(\pi)$ it is possible to find an alternating cycle decomposition $\mathcal{C}$ of $G(\pi)$ that contains at least $\lceil \frac{y}{2} \rceil$ ladder 2-cycles and $z$ independent 2-cycles.*

Theorem 25 (Christie)

*Let* $\pi \in S_n$. *Then*

$$d(\pi) \leq b(\pi) - \frac{1}{2}c_2(\pi).$$

Proof.

Using theorem 24, first find an alternating cycle decomposition $\mathcal{C}$ of $G(\pi)$ with at least $\lceil \frac{y}{2} \rceil$ 2-cycles as part of ladders and $z$ independent 2-cycles.

Proof (cont.)

- Let $k$ be the number of 2-cycles in oriented connected components of $R(\mathcal{C})$.
- Let $u$ be the number of unoriented connected components in $R(\mathcal{C})$ that include $l$ selected 2-cycles and that contain vertices representing remaining unselected 2-cycles.
- Let $v$ be the number of remaining unoriented connected components consisting only of vertices representing $m$ independent selected 2-cycles.

By theorem 21, we can sort $\pi$ using at least $k + l + u + m$ 2-reversals and only $u + v$ 0-reversals. Therefore

$$\begin{aligned} d(\pi) &\leq b(\pi) - k - l - u - m + u + v \\ &= b(\pi) - k - l - m + v \end{aligned}$$

Left to show: $-k - l - m + v \leq -\frac{1}{2}c_2(\pi)$.

## Proof (cont.)

Left to show: $k + l + m - v \geq \frac{1}{2}c_2(\pi)$. We know that

1. $k + l + m \geq \left\lceil \frac{y}{2} \right\rceil + z$ as $\left\lceil \frac{y}{2} \right\rceil + z$ is the number of selected 2-cycles in $\mathcal{C}$;
2. $v \leq \left\lfloor \frac{z}{2} \right\rfloor$ as every unoriented component representing a 2-cycle represents at least one more alternating cycle (lemma 18); and
3. $|M| = y + z \geq c_2(\pi)$.

$$\begin{aligned} k + l + m - v &\geq \left\lceil \frac{y}{2} \right\rceil + z - v && (1) \\ &\geq \left\lceil \frac{y}{2} \right\rceil + z - \left\lfloor \frac{z}{2} \right\rfloor && (2) \\ &= \left\lceil \frac{y}{2} \right\rceil + \left\lceil \frac{z}{2} \right\rceil \\ &\geq \frac{1}{2}c_2(\pi) && (3) \end{aligned}$$

□

Run time: $O(n^4)$, can be improved to $O(n^2)$ (Kaplan et al.).

## Summary

- the number of alternating cycles in a breakpoint graph $G(\pi)$ is related to $d(\pi)$
- a sorting sequence of reversals can be constructed from an alternating cycle decomposition of $G(\pi)$

## Outlook

- there exists a 1.375-approximation (Berman et al.)
- MIN-SBR for signed permutations is in $P$ (Hannenhalli et al.)

# References I


[HP95] Sridhar Hannenhalli and Pavel Pevzner. "Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals". In: *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*. STOC '95. 1995, pp. 178–189. DOI: 10.1145/225058.225112.

[KS95] J Kececioglu and D Sankoff. "Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement". In: *Algorithmica* 13.1 (1995), p. 180. DOI: 10.1007/BF01188586.

[BP96] Vineet Bafna and Pavel A Pevzner. "Genome Rearrangements and Sorting by Reversals". In: *SIAM J. Comput.* 25.2 (1996), pp. 272–289. DOI: 10.1137/S0097539793250627.

# References II


[KST97] Haim Kaplan, Ron Shamir, and Robert Tarjan. "Faster and simpler algorithm for sorting signed permutations by reversals". In: vol. 29. 1997, p. 163. DOI: 10.1137/S0097539798334207.

[Chr98] David A Christie. "A 3/2-Approximation Algorithm for Sorting by Reversals". In: *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '98. 1998, pp. 244–252. ISBN: 0898714109.

[BHK01] Piotr Berman, Sridhar Hannenhalli, and Marek Karpinski. "1.375-Approximation Algorithm for Sorting by Reversals". In: *Electronic Colloquium on Computational Complexity (ECCC)* 8 (Jan. 2001).