A 1.375-Approximation Algorithm for Sorting By Transpositions

Isaac Elias
Royal Institute of Technology

Tzvika Hartman
Weizmann Institute of Science

Sorting By Transpositions

Input: A permutation π.
Output: The least number of transpositions for sorting π.

A transposition is an operation witch switches two adjacent blocks in a permutation.

$$1\ 2\ \boxed{3\ 4\ 5}\ \boxed{6\ 7}\ 8\ 9$$
$$1\ 2\ \boxed{6\ 7}\ \boxed{3\ 4\ 5}\ 8\ 9$$

Example of Sorting

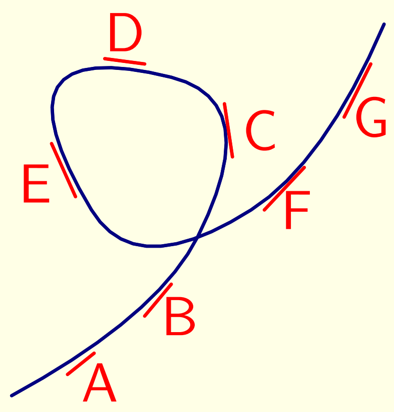
$$\boxed{5\ 4}\ \boxed{3\ 2}\ 1 \longrightarrow 3\ \boxed{2\ 5}\ \boxed{4\ 1} \longrightarrow \boxed{3\ 4}\ \boxed{1\ 2}\ 5 \longrightarrow 1\ 2\ 3\ 4\ 5$$

Genome Rearangements

Elements represent genes on a chromosome.

Segments on the chromosome can be:

- Transposed - caused by transposomes
- Reversed
- Transreversed
- Deleted
- Duplicated



$\pi =$ A B C D E F G

The Genome Rearangement Problem

Given two genomes find out what rearangment events have occured.

Bacteriophage - Transposome

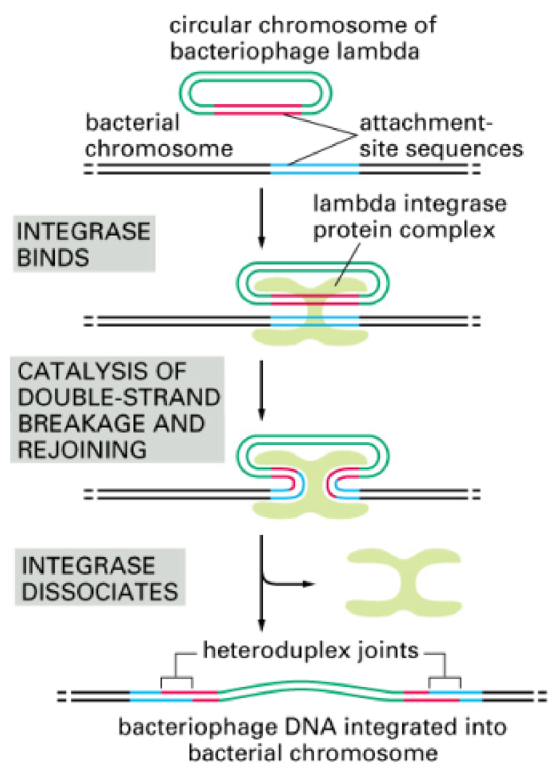


Figure 5–80. Molecular Biology of the Cell, 4th Edition.

Bacteria inserts a movable segement - a transposome.

Why study GR?

Evolution

- Rare events - allow for phylogenetic inference further back
- Large scale data: takes the whole genome into consideration
- Better multi-species analysis

Cancer

- Cancer cells undergo many genome rearangments.
- Used for cancer research, distinguish between benign and malignant tumors, diagnostics, etc.

Previous Results

SBT	1.5-approx [Bafna Pevzner, Christie, Hartman]	NP/P ?
Transposition Diameter (longest distance)	$\leq \frac{2n}{3}$ [Erikson et.al.]	$\geq \lfloor \frac{n+1}{2} \rfloor + 1$ [Christie, Meidanis et.al.]

Our Results

SBT	1.375-approx	
Transposition Diameter		$\geq \lfloor \frac{n+2}{2} \rfloor + 1$

Diameter for:

Simple permutations	$\lfloor n/2 \rfloor$
2-permutations	$n/2$
3-permutations	$\lesssim \frac{11n}{24}$

The Breakpoint Graph [Bafna Pevzner]

1 2 3 4 5

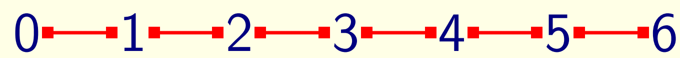
The Breakpoint Graph [Bafna Pevzner]

1. Add 0 and $n+1$ to the beginning and end and give each element a left and a right vertex.

0▪ ▪1▪ ▪2▪ ▪3▪ ▪4▪ ▪5▪ ▪6

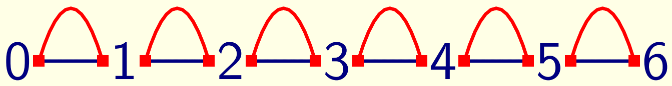
The Breakpoint Graph [Bafna Pevzner]

1. Add 0 and $n+1$ to the beginning and end and give each element a left and a right vertex.
2. Connect adjacent elements with an edge.



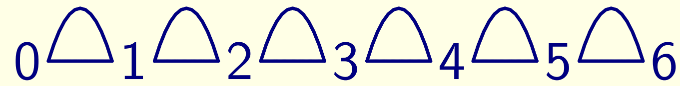
The Breakpoint Graph [Bafna Pevzner]

1. Add 0 and $n+1$ to the beginning and end and give each element a left and a right vertex.
2. Connect adjacent elements with an edge.
3. Connect successive elements by an arc.

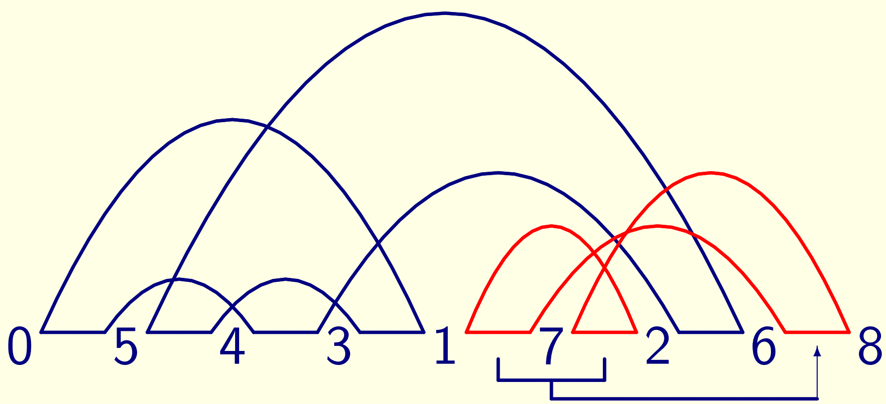


The Breakpoint Graph [Bafna Pevzner]

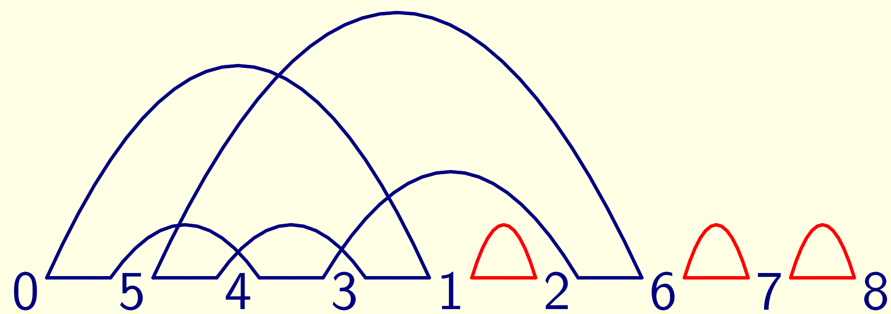
1. Add 0 and $n+1$ to the beginning and end and give each element a left and a right vertex.
2. Connect adjacent elements with an edge.
3. Connect successive elements by an arc.



Decomposes into cycles. Length of cycle = Number of arcs.



One 5-cycle and one 3-cycle.



One 5-cycle and three 1-cycles.

A transposition cuts 3 edges.

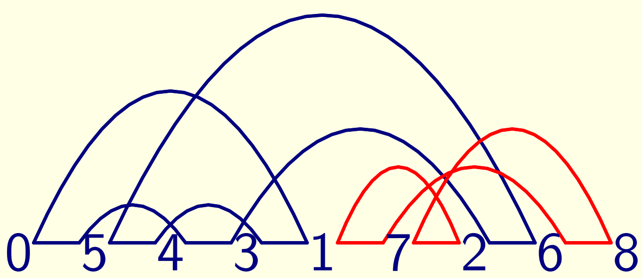
A Lower Bound [Bafna Pevzner]

Game Create $n+1$ odd cycles in as few moves as possible.

k-move the number of odd cycles is increased by k cycles.

Lemma There are only 2, 0, and -2 moves.

Lower bound $$\mathbf{d}(\pi) \geq \frac{\mathbf{n+1-c_{odd}}(\pi)}{\mathbf{2}}$$



$$d(\pi) \geq \frac{8-2}{2} = 3$$

Making Approximation Algorithms

Do not use -2-moves!

Notation (x, y): Sequence of x moves with y 2-moves.

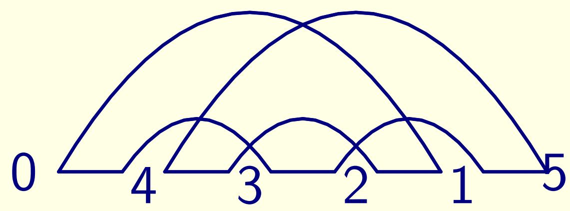
Lemma [BP] There is always $(3, 2)$-sequence; for every three moves at least two 2-moves can be performed.

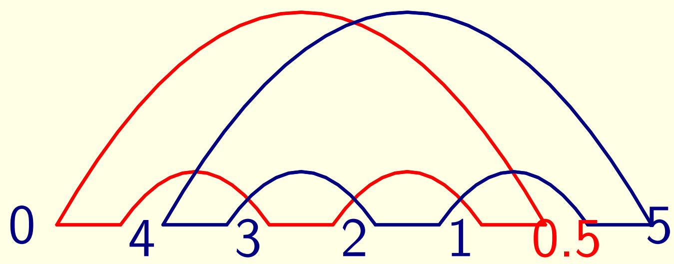
$$\Rightarrow \quad \frac{3}{2} = 1.5\text{-approximation}$$

Our algorithm uses $(11, 8)$-sequences; for each 11 moves it uses at least 8 2-moves.

$$\Rightarrow \quad \frac{11}{8} = 1.375\text{-approxmation}$$

Adding Elements - Simple Permutations [Lin Xue]





The 5-cycles is broken into two 3-cycles.

All cycles can be broken down into cycles of length ≤ 3, called simple permutations.

Elements can be added without changing the lower bound.

$$\frac{n+1-c_{odd}(\pi)}{2} = \frac{n'+1-c_{odd}(\pi')}{2} \quad \Rightarrow \quad \frac{4+1-1}{2} = \frac{5+1-2}{2}$$

A 1.375-Approximation

Step 1 Simplify π

$$\begin{array}{ccccc} \pi & \rightarrow & \pi' \ldots & \rightarrow & \pi^{(k)} \\ d(\pi) & \leq & \ldots & \leq & d(\pi^{(k)}) \\ \frac{n+1-c_{odd}(\pi)}{2} & = & \ldots & = & \frac{n+1+k-c_{odd}(\pi^{(k)})}{2} \end{array}$$

Step 2 Find sorting using only $(11,8)$-sequences for $\pi^{(k)}$

Step 3 Use sorting of $\pi^{(k)}$ to sort π.

How to do Step 2?

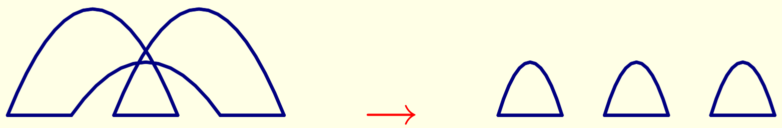
Observations

1. If find a sequence (x, y) s.t. $\frac{x}{y} \leq \frac{11}{8}$ then ok, e.g. $(1, 1)$ and $(4, 3)$.

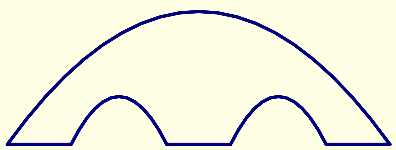
2. If there is a **2-cycle** then there is a 2-move. Ok!

3. There are two types of 3-cycles.

Oriented 3-Cycle
Has a 2-move. Ok!

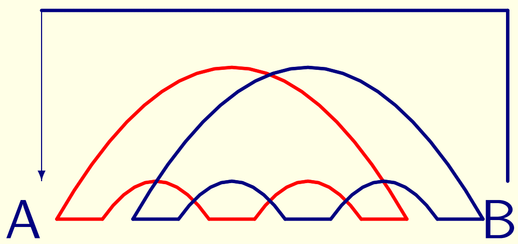


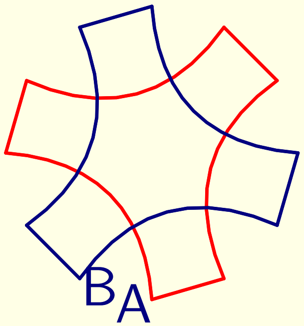
Unoriented 3-Cycle
Does not have 2-move. Not ok!



$\Rightarrow$ **Only unoriented 3-cycles!**

4. [Hart] Sorting linear permutations ⇔ Sorting circular permutations
Relative structure of the cycles matters (cyclical shift, mirroring).

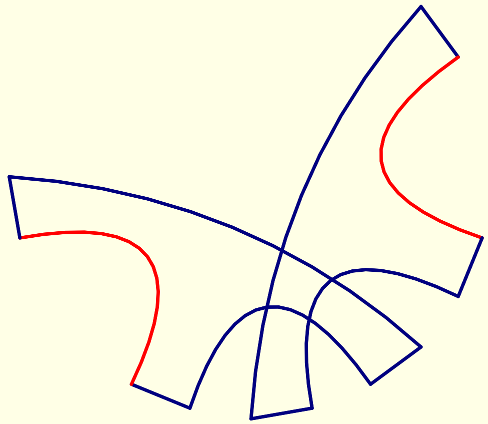




⇒ **Analyse structures of unoriented 3-cycles.**

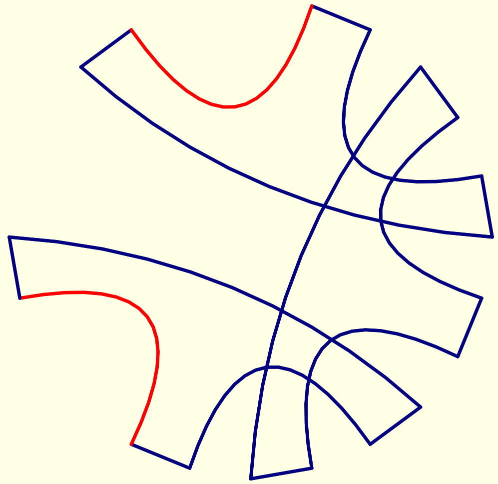
Configurations of Unoriented 3-cycles

Lemma [BP] In a breakpoint graph every arc has to cross another arc.



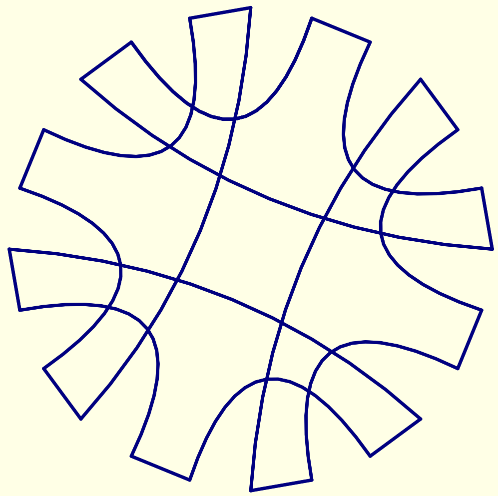
Configurations of Unoriented 3-cycles

Lemma [BP] In a breakpoint graph every arc has to cross another arc.



Configurations of Unoriented 3-cycles

Lemma [BP] In a breakpoint graph every arc has to cross another arc.

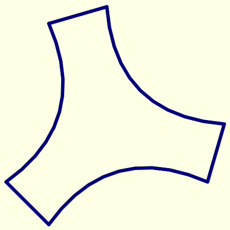


Configurations can be built by adding cycles intersecting with another cycle.

Idea A program that analyses configurations to see if an $\frac{11}{8}$-seq always exists.

Breadth First Search to Prove Existence of $\frac{11}{8}$-seq

1. Initiate a queue to contain the configuration with one cycle.



2. While the queue is non-empty do:

(a) Remove the first configuration, A, from the queue.

(b) For each way of adding a cycle; B extension A:

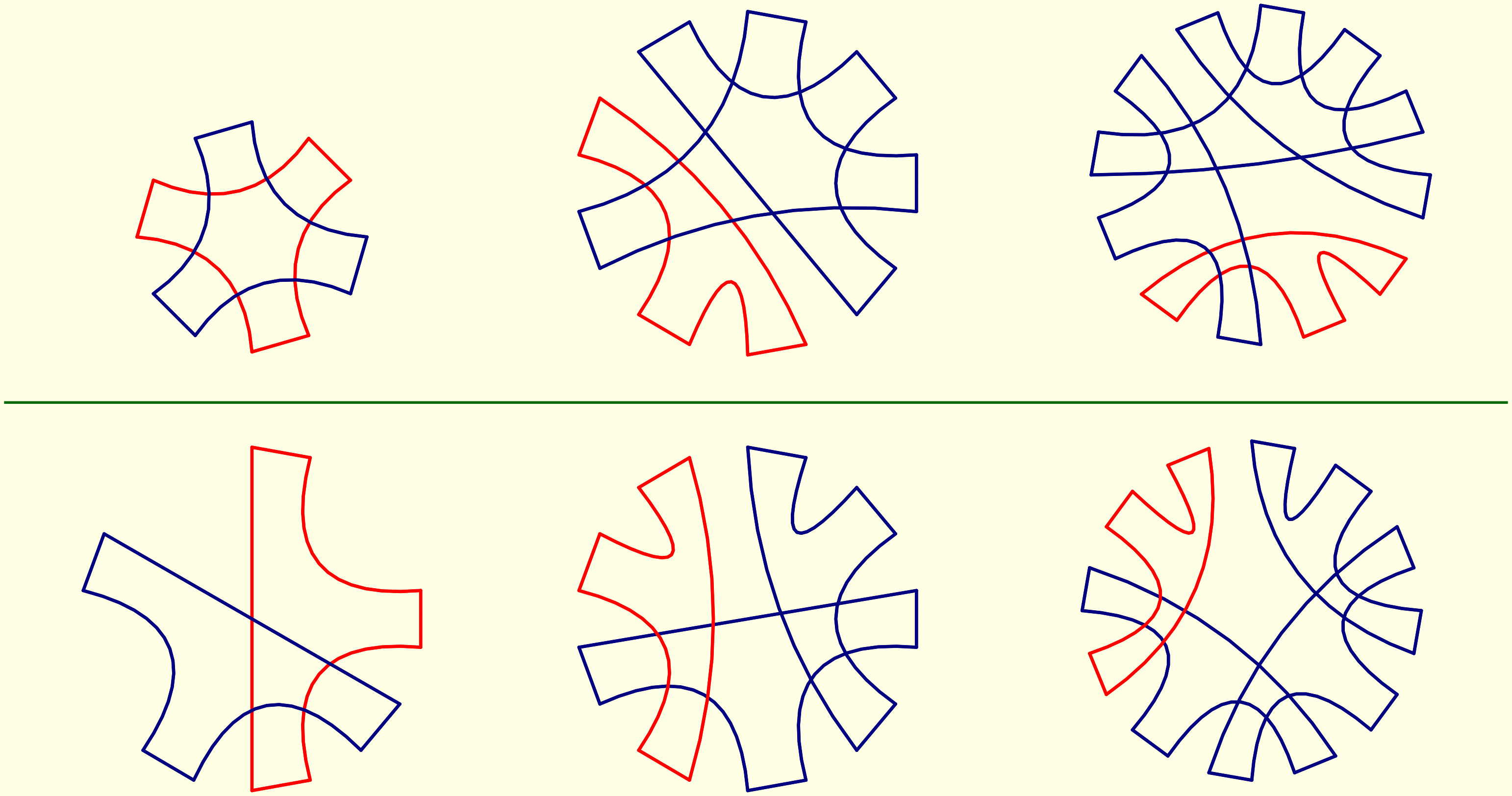
i. If B has $\frac{11}{8}$-seq then all permutations containing B has one too.

ii. Otherwise add B to the queue and continue analyzing it in next iteration.

Has to stop otherwise $\frac{11}{8}$ doesn't exist!

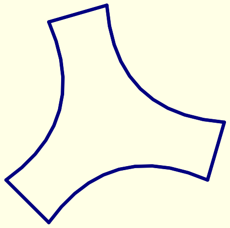
Adding Cycles to Configurations without $\frac{11}{8}$-seq

ITER 1 **ITER 2** **ITER 3**



What happens in the BFS?

INIT The queue contains one configuration.



ITER 1 For each configuration in the queue add cycles in all possible ways.

- If extension has $\frac{11}{8}$-seq then ok.
- Otherwise add to queue and continue adding cycles in ITER 2.

. . .

ITER 9 The queue is empty (all configurations of 9-cycles have $\frac{11}{8}$-seq).

Computer aided proof with 80,000 cases.

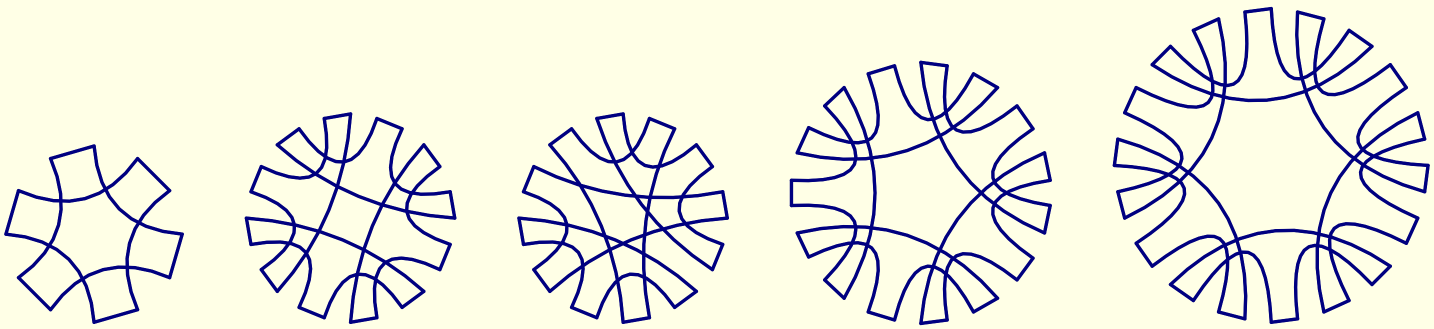
Bad Small Components

We want to show:

Lemma Every 3-permutation with ≥ 8 cycles has an $\frac{11}{8}$ sequence.

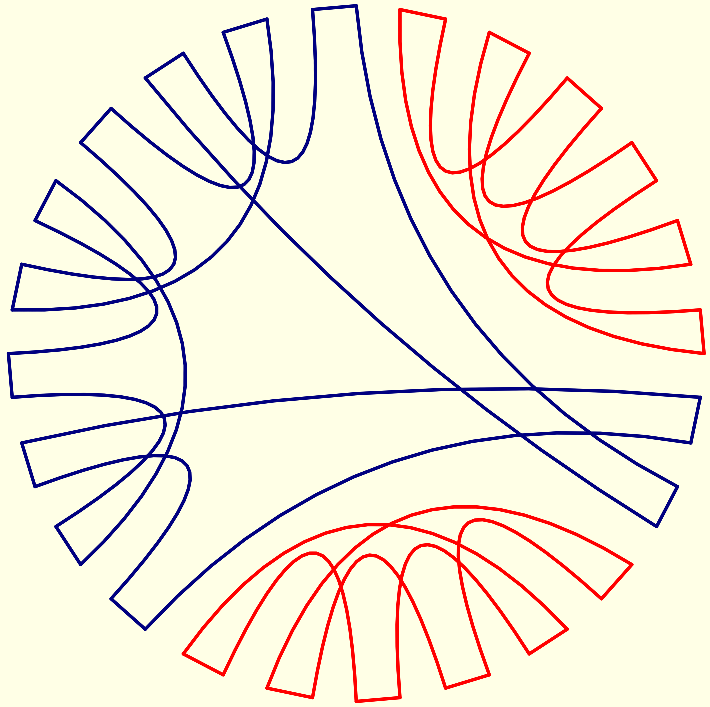
We have shown that components ≥ 9 cycles have an $\frac{11}{8}$-sequence.

Components < 9 cycles that do not have (x, y).



New case analysis showing that combinations with ≥ 8 cycles have $(11, 8)$-seq.

Example of Combination



The Approximation Algorithm

Algorithm $Sort$ **(π)**

1. Transform permutation π into a simple permutation $\hat{\pi}$.

2. Check if there is a $(2,2)$-sequence. If so, apply it.

3. While $G(\hat{\pi})$ contains a 2-cycle, apply a 2-move.

4. While $G(\hat{\pi})$ contains at least 8 cycles apply a $(4,3)$ or an $(11,8)$ sequence.

5. While $G(\hat{\pi})$ contains a 3-cycle, apply a $(3,2)$ sequence.

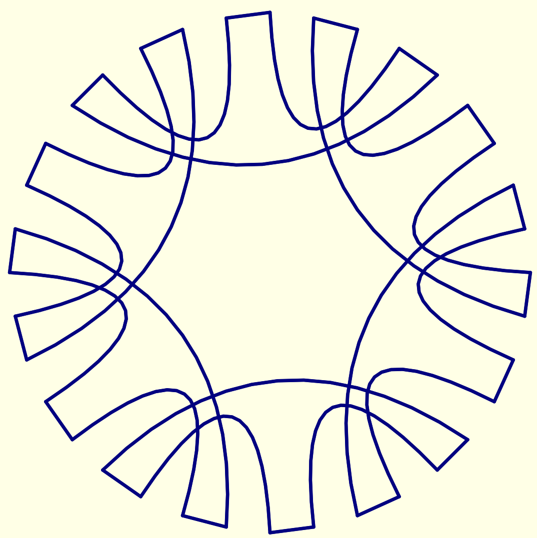
6. Mimic the sorting of π using the sorting of $\hat{\pi}$.

Can we do better?

If we analyse even bigger components can we do better?

Probably not! It seems as if the unoriented necklace can not be sorted better than with $(11, 8)$-sequences.

A new lower bound is probably needed!



Diameter for 3-permutations

Definition The longest sorting distance for any permutation made up only of 3-cycles.

Today: Every 3-permutation can be sorted with $(11, 8)$ sequences.

If a permutation has k 3-cycles it can be sorted using

$$\sim \frac{k}{8} \cdot 11 \text{ moves.}$$

All cycles length 3 so $n = 3 \cdot k$.

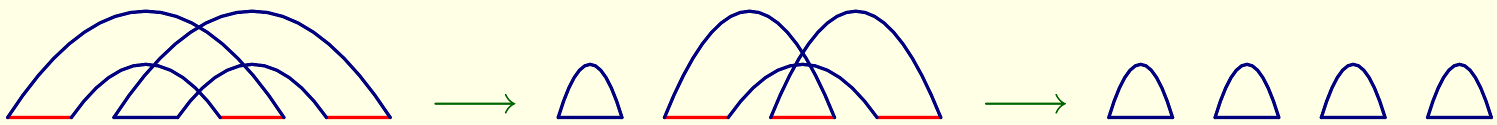
Hence upper bound

$$\lesssim \frac{11k}{8} = \frac{11n}{24}$$

Diameter for 2-permutations

Definition The longest sorting distance for any permutation made up only of 2-cycles.

All arcs must cross other arc.



Creates 4 1-cycles in 2 moves.

The identity has $n+1$ cycles.

Every 2-permutation is sorted using

$$= \frac{n+1}{4} \cdot 2 = \frac{n+1}{2} \text{ moves.}$$

Diameter for Simple Permutations

Definition The longest sorting distance for any permutation made up only of 2-cycles and 3-cycless.

Same kind of proof as for 2-permutations.

About 10 cases to analyse.

Same diameter as for 2-permutation:

$$\sim \frac{n+1}{2} \text{ moves.}$$

General Diameter

Definition The longest sorting distance for any permutation.

Earlier conjecture the reversed permutation hardest to sorted.

We show that:

$\pi =$ 0 4 3 2 1 5 13 12 11 10 9 8 7 6 14   2-permutation   n+1

requires one more move than the reversed to be sorted.

Reversed require $\sim \frac{n}{2} + 1$ moves and π requires $\frac{n}{2} + 2$ moves.

Many interesting open questions.

Cycles seem to partion into groups and combinations of these groups are hard to sort.

Results

SBT	1.375-approx	
Transposition Diameter		$\geq \lfloor \frac{n+2}{2} \rfloor + 1$
Diameter for:		
Simple permutations	$\lfloor n/2 \rfloor$	
2-permutations	$n/2$	only 2-cycles
3-permutations	$\lesssim \frac{11n}{24}$	only 3-cycles

Diameters for circular permutations.

Acknowledgments

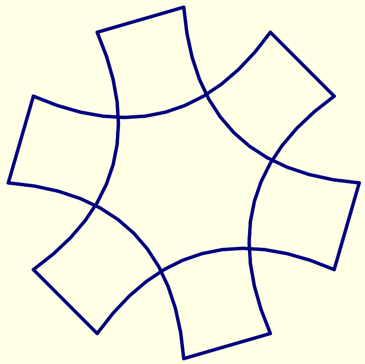
Our advisors
Prof. Jens Lagergren
and
Prof. Ron Shamir

Elad Verbin

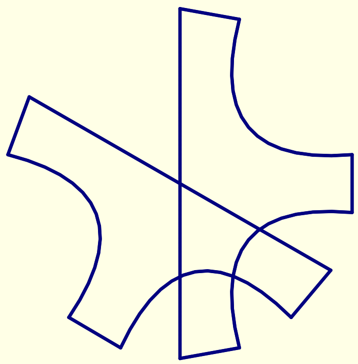
Thanks!

A $(3,2)$-sequence = 1.5 Approximation

There are two configurations with two cycles:

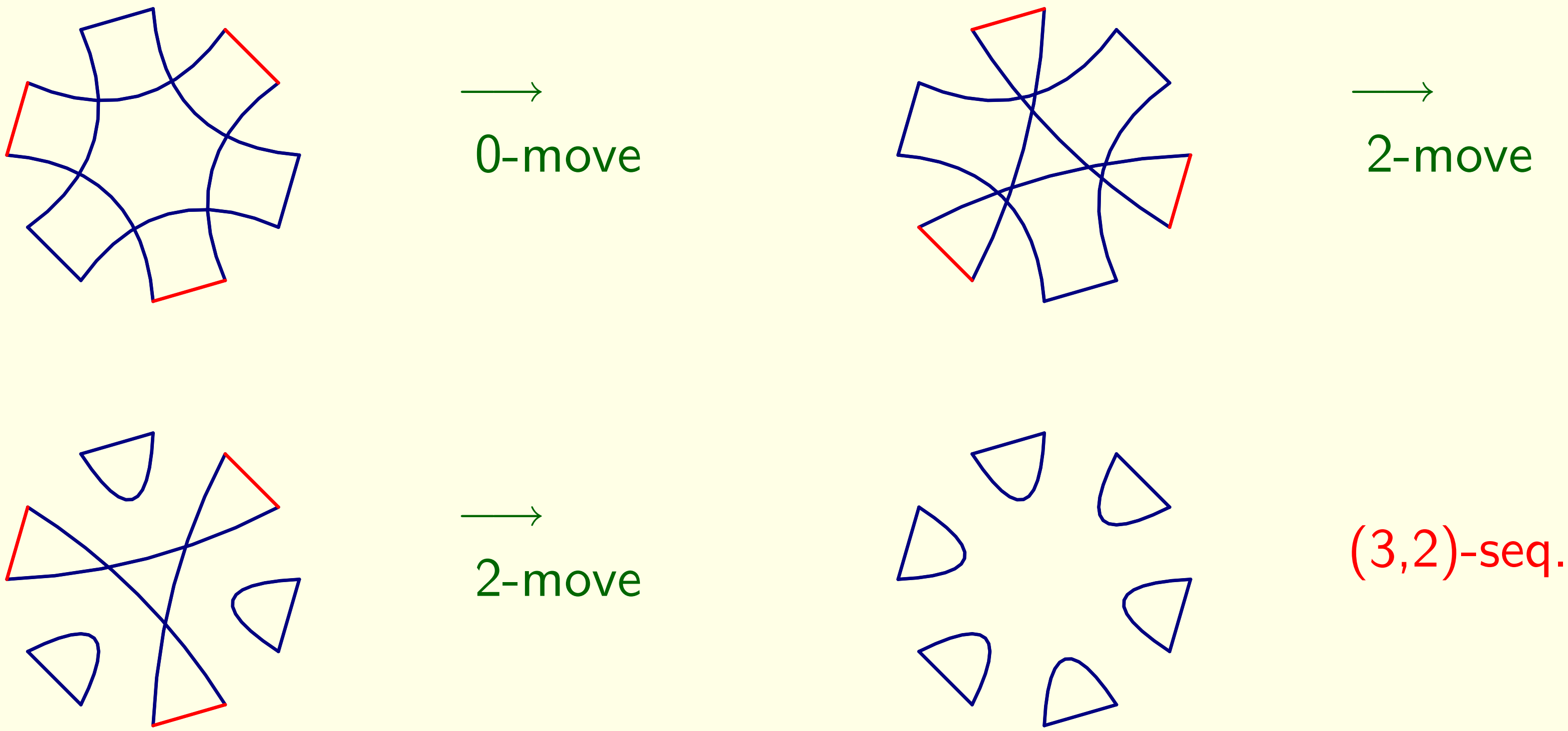


Interleaving cycles
$(3,2)$-sequence exist!



Two intersecting cycles
$(3,2)$-sequence does not exist.

Sorting Two Interleaving

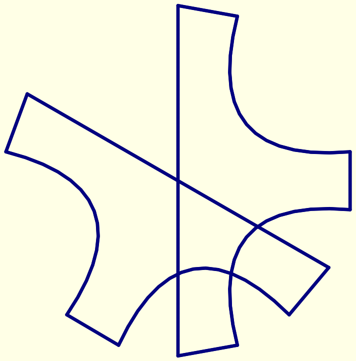


A $(3,2)$-sequence = 1.5 Approximation (cont.)

There are two configurations with two cycles:

Interleaving cycles

$(3,2)$-sequence exist!



Two intersecting cycles

$(3,2)$-sequence does not exist.

But every extension has a $(3,2)$-sequence.

$\Rightarrow$ There is always a $(3,2)$-sequence.

Sorting Extension of Two Intersecting