Combinatorics on Words

Sergey Kitaev

University of Strathclyde

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)
- Topology (knots and train tracks)

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)
- Topology (knots and train tracks)
- Number Theory (transcendental number theory)

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)
- Topology (knots and train tracks)
- Number Theory (transcendental number theory)
- Dynamical Systems/Ergodic Dynamics/Symbolic Dynamics (sofic shifts; shifts of finite type)

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)
- Topology (knots and train tracks)
- Number Theory (transcendental number theory)
- Dynamical Systems/Ergodic Dynamics/Symbolic Dynamics (sofic shifts; shifts of finite type)
- Discrete Geometry

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)
- Topology (knots and train tracks)
- Number Theory (transcendental number theory)
- Dynamical Systems/Ergodic Dynamics/Symbolic Dynamics (sofic shifts; shifts of finite type)
- Discrete Geometry
- Biology (DNA sequencing; molecular biology; study of stem cells)

Words are everywhere!

Mathematicians are in the business of producing strings of symbols (words). In some parts of maths, strings of symbols themselves become the objects of mathematical attention. Combinatorics on words (CoW) deals with combinatorial problems arising from such strings, which come up in

- Mathematical Logic
- Computer Science (data compression; cryptology; algorithms)
- Automata and Formal Languages
- Algebra (word equations in free groups; semiugroups)
- Topology (knots and train tracks)
- Number Theory (transcendental number theory)
- Dynamical Systems/Ergodic Dynamics/Symbolic Dynamics (sofic shifts; shifts of finite type)
- Discrete Geometry
- Biology (DNA sequencing; molecular biology; study of stem cells)
- Theoretical Physics (quasicrystal modelling; aperiodic structures)

The origins of CoW

- CoW was originated in the **beginning of last century** with the works of the Norwegian mathematician Axel Thue (1863–1922) in which he investigated repetitions in words.

The origins of CoW

- CoW was originated in the **beginning of last century** with the works of the Norwegian mathematician Axel Thue (1863–1922) in which he investigated repetitions in words.
- A few years later in the **1920's**, Marston Morse and Gustav Hedlund were pioneering a new branch of mathematics known as Symbolic Dynamics, inspired by the study of various classical dynamical systems dating back to Newton.

The origins of CoW

- CoW was originated in the **beginning of last century** with the works of the Norwegian mathematician Axel Thue (1863–1922) in which he investigated repetitions in words.
- A few years later in the **1920's**, Marston Morse and Gustav Hedlund were pioneering a new branch of mathematics known as Symbolic Dynamics, inspired by the study of various classical dynamical systems dating back to Newton.
- During the **latter half of the twentieth century**, CoW was developing simultaneously in different areas of mathematics including Symbolic Dynamics, Algebra, and Topology.

The origins of CoW

- CoW was originated in the **beginning of last century** with the works of the Norwegian mathematician Axel Thue (1863–1922) in which he investigated repetitions in words.
- A few years later in the **1920's**, Marston Morse and Gustav Hedlund were pioneering a new branch of mathematics known as Symbolic Dynamics, inspired by the study of various classical dynamical systems dating back to Newton.
- During the **latter half of the twentieth century**, CoW was developing simultaneously in different areas of mathematics including Symbolic Dynamics, Algebra, and Topology.
- Today, CoW has evolved into a diverse and active field of Discrete Mathematics with connections to many different areas of Mathematics, Computer Science, Biology and Physics.

The origins of CoW

- CoW was originated in the **beginning of last century** with the works of the Norwegian mathematician Axel Thue (1863–1922) in which he investigated repetitions in words.
- A few years later in the **1920's**, Marston Morse and Gustav Hedlund were pioneering a new branch of mathematics known as Symbolic Dynamics, inspired by the study of various classical dynamical systems dating back to Newton.
- During the **latter half of the twentieth century**, CoW was developing simultaneously in different areas of mathematics including Symbolic Dynamics, Algebra, and Topology.
- Today, CoW has evolved into a diverse and active field of Discrete Mathematics with connections to many different areas of Mathematics, Computer Science, Biology and Physics.
- CoW has its own Math subject classification **68R15** (under Discrete Mathematics in relation to Computer Science)

Comprehensive sources to learn about CoW

Books

- M. Lothaire. *Combinatorics on Words*. In: Encyclopedia of Mathematics and its Applications, vol. 17, Addison-Wesley, Reading, Mass., 1983.
- M. Lothaire. *Combinatorics on Words*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1997. Corrected reprint of the 1983 original.
- M. Lothaire. *Algebraic Combinatorics on Words*. In: Encyclopedia of Mathematics and its Applications, vol. 90, Cambridge University Press, Cambridge, 2002.
- M. Lothaire. *Applied Combinatorics on Words*. Cambridge University Press, 2005.

where "M. Lothaire" is a pseudonym representing several authors.

Example 1: CoW in Algebra

Semigroups

A semigroup is a set S of elements a, b, $c, \ldots$ in which an associative operation $\bullet$ is defined. The element z is a zero element if $z \bullet a = a \bullet z = z$ for all a in S.

Example 1: CoW in Algebra

Semigroups

A semigroup is a set S of elements a, b, $c, \ldots$ in which an associative operation $\bullet$ is defined. The element z is a zero element if $z \bullet a = a \bullet z = z$ for all a in S.

A Burnside type question

Let S be a semigroup generated by three elements, such that the square of every element in S is zero (thus, $a \bullet a = z$ for all a in S). Does S have an infinite number of elements?

Example 1: CoW in Algebra

Semigroups

A semigroup is a set S of elements a, b, $c, \ldots$ in which an associative operation $\bullet$ is defined. The element z is a zero element if $z \bullet a = a \bullet z = z$ for all a in S.

A Burnside type question

Let S be a semigroup generated by three elements, such that the square of every element in S is zero (thus, $a \bullet a = z$ for all a in S). Does S have an infinite number of elements?

Answering the question

Yes, it does! Thue (1906), Arshon (1937), Morse (1938). E.g. one solution is **iterating the morphism** $a \to abc$, $b \to ac$, $c \to b$: $a \to abc \to abcacb \to abcacbabcbac \to \cdots$ (square-free words)

Example 2: Symbolic Dynamics

Dynamical systems

A dynamical system is a pair (X, T) where $T : X \to X$ (X is usually compact). We look at orbits $O(x) = \{T^n(x) | n \in \mathbb{N}\}$ and questions like "Is $O(x)$ dense in X"? **Finite partition** of $X \Rightarrow$ **finite alphabet**.

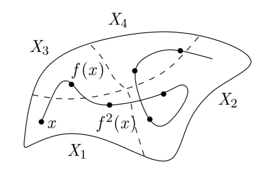
Example 2: Symbolic Dynamics

Dynamical systems

A dynamical system is a pair (X, T) where $T : X \to X$ (X is usually compact). We look at orbits $O(x) = \{T^n(x) | n \in \mathbb{N}\}$ and questions like "Is $O(x)$ dense in X"? **Finite partition** of $X \Rightarrow$ **finite alphabet**.

Encoding of the orbit of an element

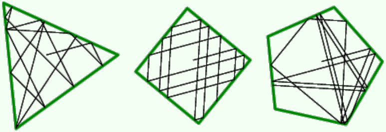
A trajectory of x in space $X = X_1 \cup X_2 \cup X_3 \cup X_4$:



The orbit of x under f can be coded as $1312244\cdots$.

Example 2: Symbolic Dynamics

Billiards



We code trajectories by considering the sequence of touched sides.

Example 3: Number Theory

Integers

Integers written in base b are words over the alphabet $\{0, \ldots, b-1\}$. For example,

$$[171]_3 = 20100$$

$$[112347]_4 = 123123123$$

Reals

Real numbers written in base b are infinite words over the alphabet $\{0, 1, \ldots, b-1\}$. For example,

$$\pi - 3 = 0.14159265358979323846264338327950288419716 9\ldots$$

$$[\pi - 3]_2 = 0.1001000011111101101010100010001000010110 10$$

Example 3: Number Theory

Folklore result

A real number is rational **iff** its base-b expansion is **eventuially periodic**.

Example 3: Number Theory

Folklore result

A real number is rational **iff** its base-b expansion is **eventuially periodic**.

Some definitions

A complex number α is algebraic if there exists $P \in \mathbb{Q}[z]$ such that $P(\alpha) = 0$. Otherwise, it is transcendental. A number x is a Liouville number, if $\forall n$, $\exists p, q \in N$, $q > 1$, such that $0 < |x - p/q| < 1/q^n$. **Every** Liouville number is **transcendental**.

Mahler's result

In 1937 K. Mahler proved that $0.1234567891011\cdots$ is transcendental but is **not** a Liouville number.

Example 4: Theoretical Computer Science

Countable and uncountable sets

A set is countable if either it is **finite** or it is in **one-to-one correspondence** with the set of natural numbers $\mathbb{N}$. A set is uncountable otherwise.

Example 4: Theoretical Computer Science

Countable and uncountable sets

A set is countable if either it is **finite** or it is in **one-to-one correspondence** with the set of natural numbers $\mathbb{N}$. A set is uncountable otherwise.

Algorithmically unsolvable problems

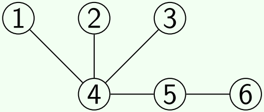
Two key ingredients in proving that there exist **algorithmically unsolvable problems** are

- showing that the **set of all words** over a finite alphabet is **countable**; and
- showing that the **set of all infinite sequences** over a binary alphabet is **uncountable**.

Example 5: Applications to Graph Theory

Prüfer codes (sequences) to encode labelled trees (1918)

Provides a proof of **Cayley's formula** (n^{n-2}) to enumerate labelled trees on n vertices.



Remove the leaf with the **smallest label** and record its neighbour:
4445 (the last neighbour does not need to be recorded)

Digital door locks

Imaging a lock with a 4 digit code (order matters!) that **remembers last three digits pressed**:

If you do not know the code (or forgot it), what is the **minimal number of presses** you need to make to open the lock (for sure)?

Example 6: Applications to Burglary

Digital door locks

Imaging a lock with a 4 digit code (order matters!) that **remembers last three digits pressed**:

If you do not know the code (or forgot it), what is the **minimal number of presses** you need to make to open the lock (for sure)?

Solving the problem

There are $10^4 = 10000$ possible codes. If one is unlucky, (s)he needs $4 \cdot 10000 =$ **40000 presses**. Using de Bruijn sequences (to be considered in the module) one needs at most **10003 presses**.

Module content

Lectures 1–5

Basic definitions in CoW, factor complexity, morphisms, fractal like sequences, primitive words, Lyndon words, periods of words, ultimately periodic sequences, powers, abelian powers, sesquipowers, Zimin algorithm, avoidability on a fixed alphabet, avoiding prohibited factors, non-transitive games, universal cycles and universal words for combinatorial structures, de Bruijn sequences, Martin's algorithm

Lectures 5–10

Universal cycles for permutations, universal partial words/cycles, crucial and bicrucial words w.r.t. forbidden sets, crucial and bicrucial permutations w.r.t. forbidden sets, word-representable graph, k-representability, graph's representation number, graphs with high representation number, semi-transitive orientation, 12-representable graphs, u-representation of graphs, k-11-representable graphs

Module logistics

- **10 Lectures** on Thursdays at 11:15–12:45

Module logistics

- **10 Lectures** on Thursdays at 11:15–12:45
- The lecture materials are to be distributed prior to the respective lectures along with several problems to work on in groups or individually; solutions to these problems will be given the following week and a brief discussion is to be conducted

Module logistics

- **10 Lectures** on Thursdays at 11:15–12:45
- The lecture materials are to be distributed prior to the respective lectures along with several problems to work on in groups or individually; solutions to these problems will be given the following week and a brief discussion is to be conducted
- Assessment is via **two homework assignments** (50% each of the final mark for the course)

Module logistics

- **10 Lectures** on Thursdays at 11:15–12:45
- The lecture materials are to be distributed prior to the respective lectures along with several problems to work on in groups or individually; solutions to these problems will be given the following week and a brief discussion is to be conducted
- Assessment is via **two homework assignments** (50% each of the final mark for the course)
- Each assignment will comprise of **5 problems** (one problem per lecture), and the assignments are to be given in the middle of the module and towards its end

Thank you for your attention!

I look forward to seeing many of you in the module ;)