29th London Stringology Days & London Algorithmic Workshop (LSD&LAW) for Costas 2025

Photo Credits: Peter G Weiner

King's College London
5th February 2025

Organising Committee:

Jacqueline Daykin (Co-Chair)
Zara Lim (Co-Chair)
Bill Smyth(Co-Chair)

Mai Alzamel
Amihood Amir
Maxime Crochemore
Roberto Grossi
Jan Holub
Gad Landau
Thierry Lecroq
Grigorios Loukides
Laurent Mouchard
Kunsoo Park
Tomasz Radzik
Jakub Radoszewski
Sohel Rahman
Wing-Kin Sung
Fatima Vayani
Bruce Watson

Schedule

9.20–9.40	Registration
9.40–9.45	Welcome by Tomasz Radzik
9:45–10.30	**Invited Talk** (Chair: Tomasz Radzik) Maxime Crochemore *Fast detection of specific fragments against a set of sequences*
10.30–11.00	**Session 1** (Chair: Laurent Mouchard) Francesco Pio Marino *Optimal Text Sampling through Set Cover* Robert Elsaesser *Fast Plurality Consensus in the Gossip and Population Model*
11.00–11.30	Coffee Break
11.30–12.00	**Session 2** (Chair: Jan Holub) Dominik Köppl *On Solving the Sparse Matrix Compression Problem Greedily* Lore Depuydt *Tag Arrays*
12:00–13:45	Lunch
13:45–14:30	**Invited Talk** (Chair: Roberto Grossi) Jakub Radoszewski *Quasiperiodicity in strings*
14:30–15:00	**Session 3** (Chair: Zsuzsanna Liptak) Holly Koponen *Computing LCCP Array for Partial Words in Linear Space* Mehrdad Atariani *A Review of Modelling Approaches of AI capabilities in human-centric systems*
15:00–15:30	Coffee Break
15:30–16:15	**Session 4** (Chair: Thierry Lecroq) Kunsoo Park *Pattern Matching in Graphs* Peter G Weiner *A Tale of Two papers From the Dawn of Stringology*
16:15–16:30	**Remembering Costas** (Chair: Jacqueline Daykin) Zara Lim Lorraine Ayad Fatima Vayani
16:45–17:30	Chapel Service
17:30–	Coal Hole

Talks

[Invited Talk] Fast detection of specific fragments against a set of sequences

Maxime Crochemore
Université Gustave Eiffel

We design alignment-free techniques for comparing a sequence or word, called a target, against a set of words, called a reference. A target-specific factor of a target T against a reference R is a factor w of a word in T which is not a factor of a word of R and such that any proper factor of w is a factor of a word of R. We first address the computation of the set of target-specific factors of a target T against a reference R, where T and R are finite sets of sequences. The result is the construction of an automaton accepting the set of all considered target-specific factors. The construction algorithm runs in linear time according to the size of $T \cup R$. The second result consists of the design of an algorithm to compute all the occurrences in a single sequence T of its target-specific factors against a reference R. The algorithm runs in real-time on the target sequence, independently of the number of occurrences of target-specific factors.

[Invited Talk] Quasiperiodicity in Strings

Jakub Radoszewski
University of Warsaw

Quasiperiodicity is a relaxed version of periodicity in strings. A *cover* (also called quasiperiod) of a string S is a substring of S whose occurrences cover the whole string S. A *seed* of S is a cover of a superstring of S. In this talk I will review fundamental algorithms for computing covers and seeds from the 1990s and discuss selected recent and new results on non-standard notions of quasiperiodicity like internal covers and λ-covers. In particular, I will show how the ideas behind the classic algorithms can be applied to efficiently compute substring covers in the internal setting.

Optimal Text Sampling through Set Cover

Francesco Pio Marino
University of Catania, University of Rouen

Abstract: The Character Distance Sampling (CDS) representation is a compact method for encoding the distances between consecutive occurrences of selected characters, known as pivots, in a string. A key challenge in optimizing this representation is determining a minimal set of pivot characters such that every substring of length m contains at least one pivot. We present a novel formulation of this problem as a variant of the Set Cover Problem, where substrings correspond to elements to be covered and characters act as candidate sets. While Set Cover is NP-hard, efficient approximation algorithms enable practical solutions, making this approach feasible for large-scale applications. This optimization has significant implications for pattern matching, text compression, and efficient substring search in massive datasets. In this paper, we establish the theoretical connection between CDS and the Set Cover Problem, propose algorithms for constructing optimal CDS representations, and demonstrate their effectiveness in avoiding worst-case search scenarios while minimizing memory usage.

Fast Plurality Consensus in the Gossip and Population Model

Robert Elsaesser
University of Salzburg

Abstract: We consider the plurality consensus problem for n agents. Initially, each agent has one of k opinions. Agents choose random interaction partners and revise their state according to a fixed transition function, depending on their own state and the state of the interaction partners. The goal is to reach a configuration in which all agents agree on the same opinion. If there is initially a sufficiently large bias towards some opinions one of them should win.

We consider this problem in two different communication models: in the sequential population model and the parallel gossip model. In the population model agents interact in randomly chosen pairs, one pair in each time step. The runtime is measured in parallel time (number of interactions divided by n). The gossip model assumes parallel rounds. During each round every agent is allowed to communicate with one randomly chosen agent. In this talk we focus on different variants of the so-called Undecided State Dynamics and present several upper and lower bounds on the number of states as well as the time needed for the protocols to converge.

On Solving the Sparse Matrix Compression Problem Greedily

Dominik Köppl
University of Yamanashi

Abstract: The sparse matrix compression problem asks for a one-dimensional representation of a binary $n \times \ell$ matrix, formed by an integer array of row indices and a shift function for each row, such that access to the matrix can be done in constant time by consulting the representation. It has been shown that the decision problem for finding an integer array of length $\ell + k$ or restricting the shift function up to values of k is NP-complete. In that light, a greedy algorithm has been proposed to shift the i-th row until it forms a solution with its predecessor rows. Despite this greedy algorithm being cherished for its good approximation in practice, we show that it actually exhibits an approximation ratio of $\sqrt{\ell + k}$.

Tag Arrays

Lore Depuydt
Ghent University

Abstract: For indexed pattern matching, we are typically asked to preprocess a text $T[1..n]$ such that later, given a pattern $P[1..m]$, we can report the locations of all the occurrences of P in T. Over fifty years we have developed an impressive collection of techniques and data structures for locating, but in this talk we argue that for pangenomics we should often not locate at all. After all, the exact location of a character in the concatenation of a set of genomes has essentially no biological significance. Since characters are sorted by context in the Burrows-Wheeler Transform (BWT), however, DNA bases that align to the same position in a reference tend to be grouped together in the BWT of a collection of genomes from the same species. Similarly, corresponding bases from the same species tend to be grouped in the BWT of many genomes from each of several species. If we store run-length compressed arrays of those positions or species in BWT order then, given the BWT interval of P, we can efficiently report which to which positions in the reference occurrences of P align or which species' genomes contain P — without locating!

Computing LCCP Array for Partial Words in Linear Space

Holly Koponen
McMaster University

Abstract: The current known computation of the Longest Common Compatible Prefix (LCCP) array of partial words using suffix trees requires $O(n^2)$ time and $O(n^2)$ space. Suffix arrays offer a more space-efficient alternative, requiring only linear space. Using a precomputed suffix array (SA) and the Longest Common Prefix (LCP) array, the brute-force approach to extend the LCP to LCCP has the worst-case time complexity of $\Omega(n^2)$ with linear memory usage. In this paper, we explore techniques to improve the brute-force computation of the LCCP array. Our ultimate objective is to use the LCCP array to compute MAXCOVER for indeterminate strings with the intent to focus on applications to biological sequences, such as DNA and proteins. As a first step, we present here the algorithms for partial words, later to be extended to work with indeterminate strings.

A Review of Modelling Approaches of AI capabilities in human-centric systems

Mehrdad Atariani
University of Law

Abstract: This research identifies major problems designers face when designing complex human-AI systems, such as AI opacity, the AI literacy gap, and collaborative design challenges. It synthesises the current research and presents the AI-powered service blueprint as a potential solution to these challenges as it maps the design and integration process of AI systems to ensure ethical, transparent, and user-centred outcomes.

These blueprints enable participatory design in which AI systems and stakeholders work together and can further be used to ensure compliance with ethical, technical, and legal standards. Whilst the concept requires empirical evaluation, it however largely represents a potential impact measure for collaborative design, understanding AI capabilities, and creating more understandable AI applications. This paper is a foundation for following empirical research to evaluate the use of AI-powered blueprints to enhance human-AI interaction in design.

Pattern Matching in Graphs

Kunsoo Park
Seoul National University

Abstract: Pattern matching is one of the fundamental problems in computer science, and it has been studied in various objects such as strings, trees, graphs, and hypergraphs. In this talk we present a fast algorithm for pattern matching in real-world graphs. We also explore the correspondence between string problems and graph problems.

A Tale of Two papers From the Dawn of Stringology

Peter G Weiner

Abstract: My talk will discuss unpublished 1973 work that solves a Stringology problem by using a Suffix Tree. The file transmission problem is to determine the best way to send a file A (assumed to be a linear string over a finite alphabet) from one computer to another via a transmission line, assuming that the receiving computer has access to another file B called the base file. In addition to sending the characters of A directly, we allow the transmission of a copy command which directs the receiving computer to append a specified, but variable length, substring of characters taken from the base file to the end of the file under construction. The cost of transmission is taken as the sum of the number of characters directly sent and K times the number of copy commands. An optimal derivation of A is a minimum-cost sequence of characters and copy commands which allow the receiving computer to construct the file A. We present an alogorithm for obtaining an optimal derivation. This algorithm is itself optimal up to a constant factor in that both its run time and storage requirements are linear functions of the lengths A and B. The results described in Yale Research Report #16, March 1973, will be presented along with historical comments.

Connecting to WiFi at King's

For Participants from Institutions affiliated with Eduroam

1. Find Eduroam in the list of WiFi options on your device.
2. Login using your Institution's email address e.g. k1234567@kcl.ac.uk and password.
3. You are now connected to the internet.

Guests

1. Find The Cloud on the list of WiFi options on your device.
2. Open a web browser to navigate to the landing page.
3. Select Create Account.
4. Complete registration by filling fields marked with an asterisk (*).
5. (Optional) You can register your device for easy access.
6. Click Continue to start accessing the internet.
7. You are now connected to The Cloud.

If you experience any issues, please contact the IT Service Desk
by email: 8888@kcl.ac.uk
or by telephone: 020 7848 8888
(Open 24/7)