ROYAL HOLLOWAY UNIVERSITY OF LONDON

CSPRC 2016

Abstracts

16th Computer Science Postgraduate Research Colloquium

Jamie Alnasir Diego Galeano

Royal Holloway University of London
Department of Computer Science

11th May 2016

Contents

Poster presentations

Preface

Welcome to the 16th Computer Science Postgraduate Research Colloquium (2016).

This colloquium serves as a forum for interaction between the faculty, staff and students engaged in the various different disciplines of research in Computer Science at Royal Holloway University of London.

This year's event is the result of a staff-student collaboration, and has been coordinated by the students. The work presented showcases the innovative ideas, diversity and excellence of postgraduate research that is conducted here, within the Department of Computer Science at Royal Holloway.

Last year we invited an alumnus speaker and this year we continue in this tradition. This provides a forum for past PhD students to share their experiences and provide us with feedback about how their time at Royal Holloway contributed to their working life.

2016's alumnus speaker is Dr. Daniel Karapetyan. Dr. Karapetyan is a former PhD student of our Department and is currently a Research Fellow in the ASAP Research Group, School of Computer Science, University of Nottingham.

We have organized the colloquium into three themes:

- Algorithms and Bioinformatics
- Machine Learning
- Types, Logic and Language Engineering

We hope you enjoy the day, and we thank you for taking part.

11th May 2015
Royal Holloway

Jamie
Diego

Hercules: An Apache Spark MapReduce algorithm for quantifying non-uniform gene expression

Jamie Alnasir

A cells transcriptome is defined as the sum total of all the messenger RNA molecules expressed from the genes of an organism and is highly dynamic and in a constant state of flux as a result of intra- and extra-cellular stimuli as well as disease pathology. RNA-Seq (RNA Sequencing) is a Next-generation, high-throughput sequencing technology that enables researchers in the biomedical and basic science fields to study various aspects of the transcriptome from alternative splicing isoforms, post-transcriptional modifications to mutations and gene expression. Such studies rely on a common task in transcriptomics which is identifying those transcripts whose expression abundance is altered by experimental conditions which differ between sets of samples and which typically employs complex computational methods in the quantification of expression levels for observed transcripts.

RNA-Seq and other Next-generation sequencing techniques are prone to biases that may be introduced in a number of the steps of a typical sequencing workflow as well as downstream computational methods. This issue has been covered in detail in the authors' paper entitled "Investigation into the annotation of protocol sequencing steps in the sequence read archive".

In addition to the problems posed by potential biases in the data, data generated from transcriptomic studies tends to be large and complex - Bigdata. Delaney characterised such datasets as possessing volume, velocity and variety. We utilise distributed computing and employ bigdata analytics tools routinely used in industry (Apache Spark) to address the challenge. To address part of the bias issue we have devised a methodology for quantifying and quality assessing non-uniform coverage of aligned transcriptomics reads within exons of a transcriptome given genome annotation and aligned reads data. To achieve this end Hercules applies a distributed programming paradigm (MapReduce) which is central to its implementation.

Improving Disease Gene's Ranking with Disease Similarities

Juan Caceres

From about 7600 currently known Mendelian diseases, 25% have no known causing genes (orphan diseases) and it is even uncertain whether every gene was discovered when the diseases have some known genes. Computational methods that allow gene prioritisation are proven valuable for discovering genomic basis in disease research. The approach of several state of the art methods can be classified as semi-supervised learning, upon some initial labelling based on known disease-gene associations, which allows a gene ranking according to the likelihood of the association between a disease and genes.

The present work analyses the addition of the Caniza disease similarity measure for seeding initial labelling in graph diffusion for disease gene ranking, in order to enhance the ranking of previously unknown disease genes. Initial results show that the inclusion of additional labelling, while predicting with old genomic information, can greatly enhance the ranking of recently discovered disease genes. The addition of initial labelling also allows the prediction of genes for orphan diseases, which is unfeasible for methods that rely on known disease genes to produce a gene ranking.

Examining protein family diversity in metagenomics data

Golestan Sally Radwan

Current metagenomics research focuses primarily on identifying known species within the sample and, in some cases, study their relative abundance under certain (normally artificial) stresses such as chemicals or other pollutants. This confinement to known/culturable species means that the research relies heavily on existing methods of analysis, such as 16S rRNA and different methods of assembly, and uses existing tools and algorithms developed for other -omics disciplines, sometimes slightly modified to fit a metagenomics pipeline.

This research reframes the problem and poses the question: "What happens to certain microbial communities if exposed to certain types of stress?" Of particular interest is the change in the "functional diversity", or the range of functions this community is able to perform. Crucially, this research views the community as a system not a set of individual species. This means that it's important to understand all functions performed in the system regardless of which species perform them, and regardless of whether these species are known or cultured. As a consequence, traditional analysis tools cannot be used and different approaches and methods had to be explored, which was the focus of this second year (part-time) of this project.

The bulk of this year's research has involved quite computationally-intensive analysis of data from public resources. As a first step, protein family motifs are analysed and converted into nucleotide sequences, then compared to the raw data. So far, experiments have included calculating the GC-content of samples, reverse-translating and location protein family motifs, and direct sample analysis to try and detect new protein families.

Next steps include generating own data in collaboration with other researchers and performing the same analysis, then refining the methods based on results, introducing some elements of machine learning, HMMs, etc as alternatives to the current brute force approach. Challenges such as limitations on storage and processing power will also need to be addressed, primarily via algorithm optimisation and potentially via better access to local or cloud-based infrastructure.

Two problems on edge coloured graphs

Bin Sheng

Edge coloured graph is a generalization of both undirected and directed graphs. We consider two problems on edge coloured graphs. The Chinese postman problem on undirected and directed graphs is polynomial-time solvable. We extend this result to edge-coloured multigraphs. Our result is in sharp contrast to the Chinese postman problem on mixed graphs, i.e., graphs with directed andundirected edges, for which the problem is NP-hard.

It is well-known that an undirected graph has no odd cycle if and only if it is bipartite. A less obvious, but similar result holds for directed graphs: a strongly connected digraph has no odd cycle if and only if it is bipartite. Can this result be further generalized to more general graphs such as edge-coloured graphs? We study this problem and show how to decide if there exists an odd properly coloured cycle in a given edge-coloured graph. As a by-product, we show how to detect if there is a perfect matching in a graph with even (or odd) number of edges in a given edge set.

Automation of Algorithm Design for Optimisation Problems

Dr. Daniel Karapetyan

With modern methods and computing power, we can tackle most of the real-world combinatorial optimisation problems. However, design and maintenance of optimisation algorithms is still prohibitively expensive for many applications. To make optimisation affordable, we aim at automation of algorithm design. This talk will give an introduction into modern approaches to (semi) automated algorithm design. It will also present a recent result on simple yet flexible schema proposed specifically for automated generation of powerful metaheuristics. Additionally, I will briefly discuss my academic career and share my personal experience of job hunting.

The Validity of a Design Intervention: the Role of Online Discussion Forums in Providing Access to Accurate Health Information

Jennifer Cole

"No one knows everything, everyone knows something, all knowledge resides in humanity" (Pierre Lévy, Chair of Collective Intelligence, Ottawa University).

The world wide web enables virtually all the knowledge in the world to be stored in such a way that anyone (with access to the internet) can both contribute to this knowledge and draw from it in real time. Lévy predicted, 20 years ago, that "communications technologies will... enable us to think collectively rather than simply haul masses of information around with us" but if we are to truly think collectively, so that we can turn collective knowledge into genuinely collective intelligence, we need appropriate navigation and retrieval systems: a classic Big Data dilemma.

How Collective Intelligence can be harnessed will depend on what the intelligence is needed for. The Collective Intelligence Genome (CIG) developed by Professor Thomas Malone, Chair of Collective Intelligence at MIT, categorises four key characteristics of Collective Intelligence systems that can be used as a design guide e.g. should the end result be a Collection (such as YouTube, where videoclips are collected together in one place) or a Collaboration (such as MIT's Climate Co-Lab project, which enables researchers from across the world to come up with collaborative solutions to climate change issues)?

Can Lévy's Theory of Collective Intelligence and Malone's Collective Intelligence Genome help us to understand which of the myriad technologies and systems available for crowdsourcing information are most appropriate to health information, and in particular for information seeking during public health emergencies?

Interviews with NGO workers in West Africa during the Ebola Crisis identified what information is sought during a public health emergency and what helps information seekers to trust it. Mapping this to the Collective Intelligence Genome suggests that the news aggregator and discussion forum

website Reddit (www.reddit.com) is an excellent match for the requirements, fulfilling most of the information needs of the health-seekers. The important question then becomes: is health information on reddit good quality? A key challenge for health information seekers is how to determine whether health information found on the internet is accurate, complete and unbiased.

My study has explored whether the quality of information on Reddit is high enough for it to be sufficiently trustworthy. What mechanisms maintain, or could be introduced to improve, the accuracy, completeness and therefore trustworthiness of health information online, particularly for use during a public health emergency?

Counting Casualties and Long Distance Dependencies in Sequential Data

Andrej Žukov Gregorič

We describe the Iraq Body Count Corpus (IBC-C) dataset, the first substantial armed conflict-related dataset which can be used for conflict analysis. IBC-C provides a ground-truth dataset for conflict specific named entity recognition, slot filling, and event de-duplication. IBC-C is constructed using data collected by the Iraq Body Count project which has been recording casualties resulting from the ongoing war in Iraq since 2003. We describe the dataset's creation, how it can be used for the above three tasks and provide initial baseline results for the first task (named entity recognition) using Conditional Random Fields and Recursive Neural Networks.

In contrast to usual named entity recognition, IBC3 named entities are significantly harder to recognise and often depend on other words far away. We look at the problem of long distance dependencies and summarise recent developments and possible ways forward.

Free Jazz in the Land of Algebraic Improvisation

Claudia Chiriţă

We discuss the connection between free-jazz music and service-oriented computing, and advance a method for formal, algebraic analysis of improvised performances. Through this work, we aim for a better understanding of both the creative process of music improvising and the complexity of service-oriented systems. We formalize free-jazz performances as complex dynamic systems of services, building on the idea that an improvisation can be seen as a collection of music phase spaces that organise themselves through concept blending, and emerge as the performed music. We first define music phase spaces as specifications written over a class of logics that satisfy a set of requirements that make them suitable for dealing with improvisations. Based on these specifications we then formalize free-jazz performances as service applications that evolve by requiring other music fragments to be added as service modules to the improvisation. Finally, we present a many-valued logic for specifying free jazz based on one of Anthony Braxton's graphic notations for composition notes.

Our work focuses on the musicking process itself, not on the resulting music: we do not provide a way to record the improvisation; instead, the music is played at run-time, whenever a new fragment is sublated into the performance by means of processes of service discovery, selection and binding. This opens the possibility for future developments focusing on the implementation of a specification and programming language whose operational semantics extends the logic programming of services by taking into account multiple truth values, instead of just True and False, as encountered in our formalization of free jazz.

On Extending Type Theories with Cannonical Objects with Subtyping Judgements

Georgiana Lungu

What is a correct way of extending type theories with canonical objects, such as those used by proof assistants like Coq, with subtyping judgements? One might be interested in answering this with a solution which is as close as possible to the programming model of such proof assistants and hence of practical interest. Another requirement for an extension with subtyping judgements is that the extension preserves the properties of the initial calculus. In this talk, I will present our proposal of extending such a type theory by adding subtyping judgements in signatures and briefly discuss definitionality as a means to achieve the second requirement. Another practical question is how such a calculus relates to the usual notion of subtyping. I will discuss this and finish with an application for universes subtyping.

Pattern Matching with Sequence Variables

L. Thomas van Binsbergen

The PLanCompS project has produced CBS, a language for specifying the formal semantics of software languages. CBS supports functions that can be called with an arbitrary number of arguments, a feature proven to be convenient for specifying software languages. Patterns with "sequence variables" are used to define these functions.

In this talk we motivate sequence variables with examples taken from CBS and by a comparison to Java's varargs. We introduce conventional pattern matching as it is found in many declarative and functional programming languages. We extend conventional pattern matching to allow for sequence variables, and give a simple algorithm that enables implementation.

Drug repurposing using chemical similarity

Diego Galeano

The drug discovery process takes in average 15 years and an investment of roughly US $1.8 billion for a single new drug. One of the key challenges of genomic drug discovery is to find successful relations between the chemical representation of the drugs and the genomic information. In this work we build a Drug-Drug Network based on the Tanimoto 2D Chemical Similarity (ChemSIM) between drugs using fingerprints from the SMILE representation; and then we evaluate whether this similarity can be used to predict co-localization of drugs targets in the Protein-Protein Interaction Network (PPI).

To assess whether the drug-drug similarity can be used to predict if the drugs share targets in the PPI, we calculate different metrics for the shortest path between drugs-targets in the PPI, train a binary classifier, and measure the performance using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The best AUC obtained was 0.85.

We shows that the chemical similarity can be used to predict drugs whose targets are close on the interactome. This provides an important confirmation about the importance of the chemical structure in the drug discovery process during screening of molecules, and opens the door for drug repurposing.

Multi-class Probabilistic Classification for Pairwise Coupling

Valery Manokhin

This paper studies empirically a method of turning pairwise machine-learning classification algorithms into multi-class classification predictors. We focus on techniques which allow producing multi-class probability estimates by combining output of pairwise classifiers. As two-class classification problems are much easier to solve, several authors proposed using output from two-class classification problems to obtain multi-class probabilities. Whilst some machine-learning algorithms produce pairwise class probabilities directly, many others output only class labels. For such classifiers, we first turn pairwise classification algorithms into probabilistic predictors which automatically enjoy the property of perfect calibration (validity) and are computationally efficient. As shown in Vovk, Petej and Fedorova (2016), such methods of turning machine-learning algorithms into probabilistic predictors, whilst producing imprecise (in practice almost precise for large data sets) probabilities, exhibit good performance when applied to two-class classification problems. Using algorithms covered in Vovk, Petej and Fedorova (2016), we turn a core set of machine-learning algorithms into probabilistic predictors and examine their performance in multi-class classification problems.

ExpoSE.js: Runtime Checking of Security-Critical JavaScript Programs

Duncan Mitchell

A recent increase in security conscious consumers has led to a dramatic rise in the demand for secure software, especially within web services. As the primary programming language compatible with web browsers, JavaScript has become ubiquitous - both client- and server-side - in software solutions across the world. Traditionally, testing JavaScript programs to ensure type safety or security properties has been a challenging task due to its incompatibility with static type checkers. We introduce an annotated type system for JavaScript which can describe security properties through annotations, and ExpoSE.js, a runtime type checker which uses symbolic execution to explore feasible program paths to check for violations of the annotated type system.

BabelView: Understanding Inter-language Flows in Android Apps

Claudio Rizzo

WebView is an Android component that allows applications to run an embedded web browser. In order to make WebView more powerful, the Android SDK provides a number of APIs for enabling JavaScript execution and communication between the application and the loaded web page. This opens new scenarios of attacks, posing serious threats to the end user. In particular, an attacker that takes control of the page inside the WebView may exploit all the methods exposed by the application, allowing them to steal user sensitive information.

This project considers the implications of this attacker model; we propose a way to model "maximally malicious" JavaScript executing within a given application. In particular, this allows for the detection of any data flows an attacker could achieve through the JavaScript interface and yields an estimate of the severity of possible script-injection attacks against the WebView.

SAT-solving: clause learning and emerging challenges

Gisela Rosi

It is easy to imagine many everyday situations when we need to satisfy multiple potentially conflicting constraints. It is also natural to imagine this situation in computing or engineering. In its simplest form, to which many more complex problems can be converted, the variables are Boolean valued (true/false) and propositional logic formulas can be used to express the constraints on the variables. A satisfying assignment for a formula is an assignment of the variables such that the formula evaluates to 1. Such an assignment is not guaranteed to exist. This bring us to the definition of the Boolean Satisfiability problem (SAT): Given a formula, find a satisfying assignment or prove that none exists.

SAT stands at the crossroads of logic, computer science, computer engineering, and operations research. This is why there are both theoretical and practical reasons that make this problem interesting. SAT became the first NP-complete problem (Cook, 1971). Given its theoretical hardness, the undeniable practical success of SAT has come as a surprise.

But although the success of SAT has led to its widespread industrial use. There are still enough instances that are difficult for the state-of-the-art solvers, and it is unclear if they will be able to handle the change in scale (number of variables) or nature (from new domains) of instances. Even more, many of the techniques that work in practice have left us in the dark on why they work. For example, while clause learning techniques have been very successful in contemporary industrial-level SAT-solving, it is necessary by the practitioners to adapt their clause learning heuristics to different domain specific instances. Such adaptations are usually made in an ad-hoc manner and so entail an expensive trial and error process.

The quest for better SAT solvers and a deeper theoretical understanding of SAT remains. The question we consider is how to systematically and theoretically characterise which clause learning adaptation approach to take for each class of domain specific instances. We also consider what kind of algorithm can be devised that can, on the one hand, mimic the simple backtracking technique that works so well in practice for DPLL/CDCL

algorithms and, on the other hand, be casted as a proof-search algorithm for a proof system different than resolution.

Refining Gene Ontology Annotations

Mateo Bobadilla Torres

Annotating genes into the Gene Ontology is becoming increasingly important to understand the role of genes and the relationship amongst them. This process is being done both by experiments and inferred by computational methods. Here we propose a new computational approach to annotating genes into the ontology that, in contrast with the usual computational methods, does not aim to annotate all the genes but to improve already existing annotations. We start by selecting annotations which are one level up from the leaves in the ontology. Semantic Similarity data is calculated between the selected gene and all the genes annotated in candidate terms (i.e., the potential new annotated terms). This allows us to determine the new proposed term to annotate. We integrate multiple pieces of data, which greatly improves the prediction. We find that refining already existing annotations gives a relevant improvement on the completeness of the annotations for the tested organisms.