UNIVERSITY OF
OXFORD

DEPARTMENT OF
COMPUTER
SCIENCE

Proceedings of the University of Oxford Department of Computer Science Student Conference 2014

Organising Committee: Krzysztof Bar, Gaurav Bhadra, Mateusz Tarkowski, Anthony Potter, Jan Buys, Stefano Gogioso, Amar Hadzishanovic, Anna Jones, Kylie Beattie, Luying Chen

June 2014

Foreword

The 2014 Department of Computer Science Student Conference was held on the 13th June in the department. This year we had a very decent number of submissions with 19 abstracts and 9 posters submitted. What is particularly encouraging, the 12 abstracts that were accepted represented research from across the departments research themes.

The conference is organised by students and for students and with the exceptionally high involvement of all members of the Programme Committee this year, it was a particular pleasure to organise. It was an unparalleled opportunity to develop useful skills, which will undoubtly pay dividends when we get involved in other similar scientific initiatives in the future. Some of the speakers had a chance to give their first ever academic talk and gained more confidence to give similar presentations at events outside of the department. Additionally, due to exceptionally high student involvement, all of the reviewers were themselves DPhil students, and certainly gained experience which will be useful when they review for conferences in their fields. We hope that the audience learned more about research conducted in other research groups in the department. The central concept of the student conference is to create links between different research groups in the department and make us more aware that together we form a single department.

The conference would have not been a success without the help of a significant number of people to whom the organising committee would like to express their gratitude. In particular, we would like to thank Prof. Ursula Martin, who gave the keynote talk of the student conference and judged several of the prizes. We would also like to thank prof. Stephen Pulman - the incoming Director of Graduate Studies, who helped to judge talks and selected the best poster. Special thanks go to Julie Sheppard and Wendy Adams who, as every year, helped with the preparation immensely. We are certain that the counference would not have taken place withour their continued support and dedication. Lastly, we would like to thank all of those who submitted a poster or an abstract, your willingness to contribute allowed us to present the range of research topics explored in our department.We hope that all attendees found this year's conference an enjoyable and enriching experience and would like to wish next year's programme committee the best of luck in organising the next edition.

Krzysztof Bar
Programme Committee Chair

Prizes

Best Abstract: Vardui Yeghiazaryan

Best Poster: Stefano Ortona

Best Talk: Awarded jointly to Sara Dutta and Miriam Backens

Organisation

Organising Committee

Krzysztof Bar, Gaurav Bhadra, Mateusz Tarkowski, Anthony Potter, Jan Buys, Stefano Gogioso, Amar Hadzishanovic, Anna Jones, Kylie Beattie, Luying Chen

Keynote Speaker

Prof. Ursula Martin

Special Thanks

We would like to sincerely thank Julie Sheppard and Wendy Adams for all their help, dedication and continued support for the Student Conference in the Department of Computer Science at the University of Oxford.

Also thanks to William Smith, Anthony Potter and Krzysztof Bar for chairing sessions.

Conference Programme

Contents

The Lognormal Lung

Quantifying inhomogeneity of ventilation

James Mountain

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

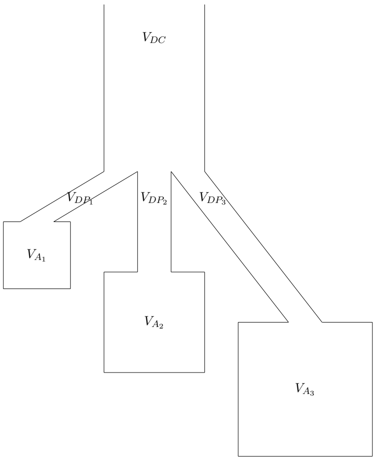


Fig. 1. An illustration of the model for the case of three alveolar compartments. V_{DC} is the common deadspace volume, V_{DPi} personal deadspace volumes and V_{Ai} alveolar volumes (volumes not to scale).

In the European Union pulmonary disease is the third most common cause of death and in North America the fourth [7]. Currently the two most common forms of assessment of pulmonary function are spirometry and plethysmography. In spirometry the most common parameter of interest is the maximum volume of gas a patient can expire in one second (FEV_1), while in plethysmography an estimate of the volume of gas in the thorax is made. To detect the presence of lung disease these measurements are compared to 'normal' values for patients of similar physical characteristics. However neither of these techniques is sensitive enough to detect early stage lung disease and both require a significant degree of patient co-operation to perform [5, 6, 9]. Thus there is a need to develop a sensitive method for detecting the early stages of lung disease.

One concept believed to be an extremely important indicator of respiratory health is the degree of inhomogeneity of ventilation [1, 3, 4]. This is usually considered by estimating the ventilation volume distribution of the lungs. This slightly abstract concept can be thought of as sampling small volume elements of the alveoli and determining how much fresh gas enters this volume during each breath. In perfectly homogeneous lungs each unit of volume would receive the same amount of fresh gas, however in practice this is not the case and even healthy lungs display some amount of inhomogeneity. The most common technique used to estimate these distributions is a multi-breath nitrogen washout (MBNW). The procedure involves switching a subject from breathing room air to pure oxygen and monitoring the concentration of nitrogen the patient expires as it washes out their lungs. A ventilation volume distribution compatible with this is then found by varying the fractional ventilations received by 50 compartments of fixed ventilation volume ration so that the data simulated with the model matches that taken experimentally [8].

However this technique has failed to gain widespread clinical acceptance for two reasons: (i) The poor quality of experimental data and (ii) the mathematical models used to represent this technique are inadequate. We are fortunate to be collaborating with Professor Peter Robbins (Department of Physiology, Anatomy and Genetics, University of Oxford) who has developed experimental apparatus to measure MBNWs with much higher accuracy, thus here we present a novel model (based around the concept that the lung can be modelled using an underlying ventilation volume distribution) that we hope will take advantage of this new more accurate data to provide a clinically useful measure of inhomogeneity of ventilation.

A schematic diagram of our model is displayed in Figure 1. We consider the alveoli as N well mixed compartments with differing ventilation volume ratios each with their own associated personal deadspace volume, with the whole lung being ventilated through a common deadspace volume. We assume that the ventilation volume ratios of the compartments are governed by an underlying log normal ventilation volume distribution, a distribution for which there is theoretical support [2]. By fitting the model to data from MBNW we hope to estimate the width of the distribution, which hopefully will provide an index of the inhomogeneity of ventilation.

In this talk I will discuss: (i) why I think this problem is interesting and the contribution I hope to make, (ii) our model and some of the mathemeatical and computational issues around its development.

References

[1] M. R. Becklake. “A new index of the intrapulmonary mixture of inspired air.” In: *Thorax* 7.1 (Mar. 1952), 111–6.

[2] B. S. Brook et al. “Theoretical Models for the Quantification of Lung Injury Using Ventilation and Perfusion Distributions”. In: *Computational and Mathematical Methods in Medicine* 10.2 (2009), 139–154.

[3] W. S. Fowler. “Lung function studies; uneven pulmonary ventilation in normal subjects and in patients with pulmonary disease.” In: *Journal of Applied Physiology* 2.6 (Dec. 1949), 283–99.

[4] W. S. Fowler, E. R. Cornish, and S. S. Kety. “Lung function studies. VIII. Analysis of alveolar ventilation by pulmonary N_2 clearance curves.” In: *The Journal of Clinical Investigation* 31.1 (Jan. 1952), 40–50.

[5] M. R. Miller et al. “Standardisation of spirometry.” In: *The European Respiratory Journal : Official Journal of the European Society for Clinical Respiratory Physiology* 26.2 (Aug. 2005), 319–38.

[6] R. Pellegrino et al. “Interpretative strategies for lung function tests.” en. In: *The European Respiratory Journal : Official Journal of the European Society for Clinical Respiratory Physiology* 26.5 (Nov. 2005), 948–68.

[7] N. Siafakas et al. “Optimal assessment and management of chronic obstructive pulmonary disease (COPD). The European Respiratory Society Task Force”. In: *European Respiratory Journal* 8.8 (1995), pp. 1398–1420.

[8] P. D. Wagner. “Information content of the multibreath nitrogen washout”. In: *Journal of applied physiology* 46.3 (Mar. 1979), 579–587.

[9] J. Wanger et al. “Standardisation of the measurement of lung volumes.” In: *The European Respiratory Journal : Official Journal of the European Society for Clinical Respiratory Physiology* 26.3 (Sept. 2005), 511–22.

Efficacy of anti-arrhythmic drugs during ischaemia

A computational modelling whole heart study

Sara Dutta

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

Since the first cardiac cell model was built over 50 years ago, the cardiac modelling field has played an important role in improving our knowledge of heart disease [4]. Computer simulations provide a powerful platform to dissect and analyse specific processes at a multi-scale level: from the ionic currents of individual cells up to the whole heart level. They allow rapid testing of hypotheses over a range of values and conditions. Thanks to the increase in computational power we can now simulate whole heart simulations on supercomputers using a multi-scale biophysically detailed approach. This allows us to investigate important drug safety concerns.

Sudden cardiac is a major killer in the western world and is usually preceded by abnormal heart rhythms, referred to as arrhythmias. Anti-arrhythmic drugs constitute first line therapy for patients suffering from such disturbances. Their efficacy has been shown in numerous studies and trials [3]. However, one of the main side effects of some anti-arrhythmic drugs is in fact increased arrhythmogenicity [2]. Mechanisms underlying this paradox are still not fully understood. One of the possible risk factors is ischaemia (see panel A of Figure 1). When part of the heart doesn't receive adequate blood supply, and numerous electrophysiological changes occur that increase the likelihood of developing arrhythmias [1]. Computer simulations have provided novel insights into mechanisms of ischemia-induced electrophysiological arrhythmic events and cardiotoxicity in non-ischemic hearts, by generating high spatio-temporal resolution data not accessible using experimental methods alone.

The present study extends this approach to the investigation of pharmacological action of anti-arrhythmic drugs in the diseased human ischaemic ventricles. We present a state-of-the-art human whole heart regional ischaemia model, including realistic ion channel dynamics and fibre architecture. Simulations were run using Chaste, an open source simulation software for multi-scale biological problems written in C++, on HECToR, the UK's national supercomputer.

The monodomain model is used to describe the electrical activity of the heart based on the cable theory. It is made up of a set of parabolic and elliptic partial differential equations (PDEs), which are connected, at each point in space, via the cell membrane, which is defined by a complex set of ordinary differential equations (ODEs) making up the cardiac action potential (AP) model, which describes the cell membrane kinetics through time. Ischaemic cells were assigned different electrophysiological properties based on experimental data to simulate hyperkalaemia, acidosis and hypoxia, as shown in panel B of Figure 1 [1]. Anti-arrhythmic drugs we simulated by inhibiting the activity of a specific ionic current by 30 and 50%. Arrhythmias were simulated by applying a premature beat at varying intervals through time, as shown in panels C and D of Figure 1. Therefore, effects of anti-arrhythmic drugs on arrhythmia dynamics and likelihood can be analysed.

In this study we present a novel computational model and simulation framework for the study of drug therapy on arrhythmic mechanisms in the human ischaemic heart. The use of Chaste, a state-of-the-art software package and the supercomputer HECToR were essential to run these computationally demanding simulations. This model shows that anti-arrhythmic drugs did not suppress arrhythmias during ischaemia but 30% drug block did decrease the

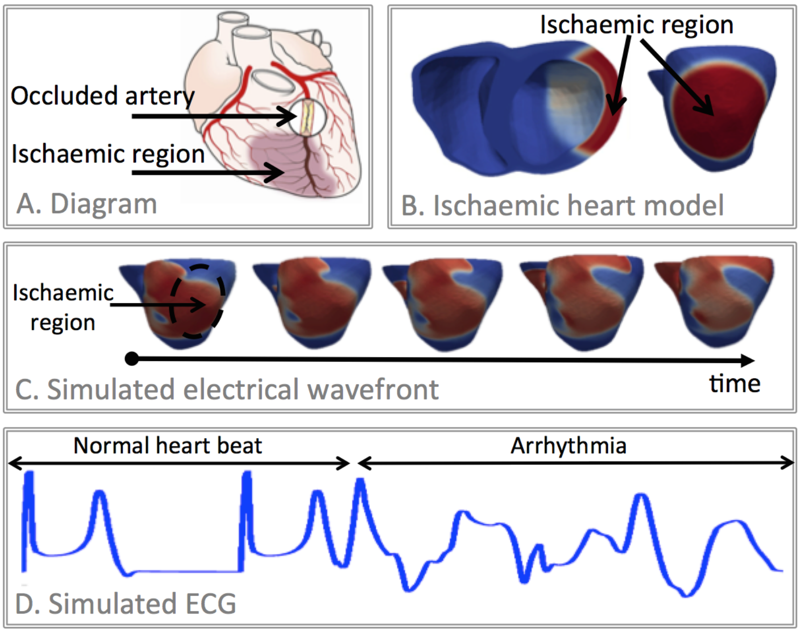


Fig. 1. Our whole heart ischaemia model, which represents the effects of an occluded artery, as shown in (A.), by simulating an ischaemic region, as shown in (B.). Simulation results include the electrical wavefront pattern, as shown in (C.), and the electrocardiogram (ECG), as shown in (D.).

likelihood of triggering an arrhythmia. The framework presented here could be extended for the investigation of other disease conditions, such as heart failure and hypertrophy, as well as other pharmacological agents targeting multiple ion channels and electrical therapy through the use of pacemakers and defibrillators.

References

1. Carmeliet, E. Cardiac ionic currents and acute ischemia: from channels to arrhythmias. *Physiological reviews 79*, 3 (July 1999), 917–1017.
2. MacNeil, D. J. The Side Effect Profile of Class III Antiarrhythmic Drugs: Focus on d,l-Sotalol. *The American Journal of Cardiology 80*, 8 (Oct. 1997), 90G–98G.
3. Wazni, O. M., Marrouche, N. F., Martin, D. O., Verma, A., Bhargava, M., Saliba, W., Bash, D., Schweikert, R., Brachmann, J., Gunther, J., Gutleben, K., Pisano, E., Potenza, D., Fanelli, R., Raviele, A., Themistoclakis, S., Rossillo, A., Bonso, A., and Natale, A. Radiofrequency ablation vs antiarrhythmic drugs as first-line treatment of symptomatic atrial fibrillation: a randomized trial. *JAMA : the journal of the American Medical Association 293*, 21 (June 2005), 2634–2640.
4. Winslow, R. L., Trayanova, N., Geman, D., and Miller, M. I. Computational medicine: translating models to clinical care. *Science translational medicine 4*, 158 (Oct. 2012).

Experiments on the Use of Fast Marching for Feature Identification

Varduhi Yeghiazaryan and Irina Voiculescu

Oxford University Department of Computer Science
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

We propose a *fully automatic* method for feature identification. Our main application is the identification of abdominal organs in computerized tomography (CT) scans. We use 3D medical data although this method is not data specific. The main steps of the method are (1) segmenting the scan data, followed by (2) applying the Fast Marching Method in order to extract feature shapes.

We segment the scan volumes using algorithms devised within the Spatial Reasoning Group here at Oxford [1, 2], which output an *image partition forest (IPF)*, a hierarchy of adjacency graphs that partition the volume (and hence each individual slice image). Significantly, the segmentation is successful at dividing the image into regions of semantic importance. However the resulting regions do not shape feature boundaries accurately: they only approximate abdominal features roughly. Thus we rely on the IPF to initialise the Fast Marching Method, but we need a post-processing step in order to detect the true boundaries.

The *Fast Marching Method* is an efficient iterative algorithm, introduced by Sethian [3, 4], for numerical approximation of the development of propagating fronts, closed hypersurfaces moving in the direction of the surface normal. Representing the front implicitly through the arrival time function $T : \mathbb{R}^n \to \mathbb{R}$ (the front reaches the point $\boldsymbol{x}$ at time $T(\boldsymbol{x})$), it solves the *eikonal* equation $|\nabla T(\boldsymbol{x})| = \frac{1}{F(\boldsymbol{x})}$ with a discretised model on a lattice, appropriate upwind schemes and optimal ordering of points in space.

Given a particular abdominal feature (such as an organ), we filter the IPF for regions based on that feature's anatomical knowledge. For instance, regions corresponding to the right kidney should be in a high layer of the IPF, have a mean greyscale value in a certain interval (known from radiology studies [5], inferred from the typical radiodensity Hounsfield Unit (HU); the right kidney should be 'west' of the spine in an axial slice, and anatomically close to the spine; also, the right kidney should span a reasonable number of voxels (depending on the number of slices in the image), etc.

Having identified a suitable region corresponding to our chosen feature, we choose points within that region which are within a fixed greyscale range (again, inferred from the feature's typical HU). We use these as *seed points* to initialise the Fast Marching algorithm.

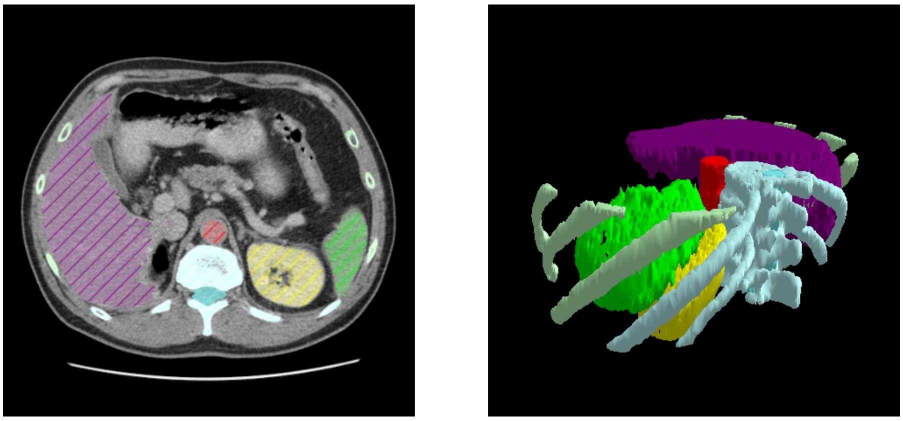
This process is repeated for each feature of interest (mainly organs, bones or blood vessels) within the abdomen.

The advancement of the Fast Marching front is governed by *speed functions* of the form

$$F(\boldsymbol{x}) = \frac{1}{\left(1 + \left(\frac{|\nabla I(\boldsymbol{x})|}{C}\right)^n\right)^m}, \quad C > 0, \;\; n, m \in \mathbb{N} \quad \text{or}$$
$$F(\boldsymbol{x}) = e^{-C|\nabla I(\boldsymbol{x})|}, \;\; C > 0$$

These were compared and tuned on CT data pre-processed with windowing or smoothing.

The development of the front in Fast Marching slows down at points with high gradient magnitude values (in particular, in the neighbourhood of organ boundaries). We captured such slow development of the front and used this as a *stopping criterion* for the Fast Marching. We finalised the identified regions with several iterations of morphological closing in



(a) Final results shown on a single axial CT slice.

(b) A 3D reconstruction of the identified features using the *marching cubes* algorithm which constructs a polygonal mesh around features which belong together.

Fig. 1. The results of 3D abdominal feature identification with Fast Marching Method. Aorta, kidneys, liver, ribs, spine, spinal cord and spleen are visible in their respective colour.

order to remove spurious holes and to smooth out the boundaries. The results of our procedure on a typical CT volume are illustrated in Figure 1.

We carried out extensive empirical experiments to identify the impact of various parameter choices on the performance of the Fast Marching Method. In particular, we determined appropriate parameter values to apply Fast Marching to 3D medical data, and we used these in our implementation, thus automating a process which was previously deemed to be carried out 'manually'.

Extensive analysis of the method's parameters allowed us to achieve significant identification results while employing a comparably simple algorithm. Results are currently validated only by a human judge. Future quantitative analysis is needed to give a more precise assessment to the performance of the method and compare it with other approaches. Also, experiments on a wider class of image volumes are needed since new medical scanners produce images with higher resolution and better precision.

References

1. Golodetz, S. *Zipping and Unzipping: The Use of Image Partition Forests in the Analysis of Abdominal CT Scans.* DPhil thesis, University of Oxford, 2011.
2. Golodetz, S. M., Nicholls, C., Voiculescu, I. D., and Cameron, S. A. Two tree-based methods for the waterfall. *Pattern Recognition* (May 2014).
3. Sethian, J. A. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences 93*, 4 (1996), 1591–1595.
4. Sethian, J. A. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 1999.
5. Tidwell, A. S. Advanced imaging concepts: A pictorial glossary of CT and MRI technology. *Clinical Techniques in Small Animal Practice 14*, 2 (1999), 65–111.

Weak and Nested Class Memory Automata

Conrad Cotton-Barratt

Oxford University Department of Computer Science

While the theory of automata over finite alphabets has been studied extensively, automata over infinite alphabets are only beginning to be investigated. Infinite alphabets are a useful abstraction in several areas of computer science, and have found applications in database theory [4] and verification [5, 6]. The infinite alphabets generally take the form $\Sigma \times \mathcal{D}$ where Σ is a finite alphabet and $\mathcal{D}$ is an infinite set of "data values", which have no underlying structure (i.e. we can only test for equality). We call languages over these alphabets "data languages".

Class Memory Automata (CMA) ([2]) are a natural kind of automata over infinite alphabets which have been used to solve previously-outstanding problems in verification, by using data values to represent names [5]. However, although CMA have a decidable emptiness problem, the complexity is high: equivalent to that of Petri Net Reachability. Further, while closed under intersection and union, they are not closed under complementation, and do not have a decidable equivalence problem.

In this work we identify a natural restriction of CMA, which we call Weak CMA. Weak CMA have significantly better complexity – emptiness is ExpSpace-complete – and in the deterministic case have closure under all Boolean operations, and thus a decidable equivalence problem. While normal CMA correspond to a very natural fragment of first order logic over infinite alphabets [3], it is less clear for weak CMA, and we are investigating this connection.

In some cases, it is useful to have more structure on the underlying data values, such as adding an ordering, or multiple levels of "nested" data. Initial work on the latter indicated that in the presence of nested data decidability was lost in the natural extension of first order logic [1]. We present a new form of automaton, Nested Data CMA, an extension of CMA to operate over multiple levels of nested data, and show that although emptiness is undecidable in general, by adding the weakness constraint decidability is recovered. Further, in the deterministic case we again get closure under all Boolean operations, and hence a decidable equivalence problem. We further show a close link between these Nested Data CMA and Higher-Order Multicounter Automata, which were introduced in [1] as a proof vehicle in their investigation of nested data.

These results are joint work with Luke Ong and Andrzej Murawski.

References

1. Björklund, H., and Bojanczyk, M. Shuffle expressions and words with nested data. In *MFCS* (2007), pp. 750–761.
2. Björklund, H., and Schwentick, T. On notions of regularity for data languages. *Theor. Comput. Sci. 411*, 4-5 (2010), 702–715.
3. Bojanczyk, M., David, C., Muscholl, A., Schwentick, T., and Segoufin, L. Two-variable logic on data words. *ACM Trans. Comput. Log. 12*, 4 (2011), 27.
4. Bojanczyk, M., Muscholl, A., Schwentick, T., and Segoufin, L. Two-variable logic on data trees and xml reasoning. *J. ACM 56*, 3 (2009).
5. Hopkins, D. *Game Semantics Based Equivalence Checking of Higher-Order Programs.* PhD thesis, Department of Computer Science, University of Oxford, 2012.
6. Segoufin, L. Automata and logics for words and trees over an infinite alphabet. In *CSL* (2006), Z. Ésik, Ed., vol. 4207 of *Lecture Notes in Computer Science*, Springer, pp. 41–57.

OpenSky

Using A Large-Scale Sensor Network For Air Traffic Research

Martin Strohmeier,
Supervised by Ivan Martinovic

Department of Computer Science
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

Due to increasing congestion of the commercial airspace more efficient air traffic management methods are required. The world's aviation authorities are currently undertaking a major upgrade from conventional air-traffic management to the Next Generation Air Transportation (NextGen) system. The Automatic Dependent Surveillance - Broadcast (ADS-B) protocol is at the heart of NextGen and is currently being rolled out in most countries. Traditional air traffic control technologies such as Primary Surveillance Radar (PSR) and Secondary Surveillance Radar (SSR) use ground-based antennas to independently measure the range and bearing of airborne objects. With ADS-B, on the other hand, aircraft determine their own position using Global Navigation Satellite Systems and broadcast it periodically over the 1090 MHz radio frequency to ground stations or other aircraft in the proximity (see Fig. 1a) for a graphical overview). Thus, one of the main advantages of ADS-B is being able to continuously broadcast exact information about altitude, heading, velocity, and other flight data, decreasing the dependence on expensive and less accurate PSR and SSR technologies. Besides lowering separation requirements between aircraft (and thus enabling more efficient higher-density airspaces), this improves the overall situational awareness of pilots and air traffic controllers significantly while reducing the costs of air traffic surveillance [3].

Whereas ADS-B strongly enhances the capabilities of air traffic surveillance systems, there are many facets of the technology that need further evaluation to ensure a quick and safe adoption. As various concerns with the ADS-B protocol such as security vulnerabilities and problems with the capacity of the wireless channel have emerged over the past years [2], it is important that these and other issues are thoroughly investigated.

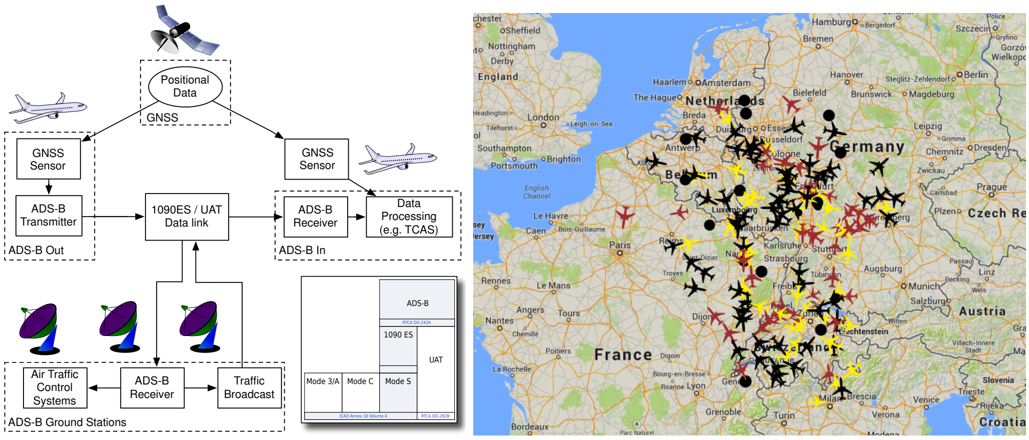


Fig. 1. a) ADS-B system architecture, including protocol hierarchy. b) Live screenshot of OpenSky reception over Central Europe.

However, until recently only closed government- and industry-related groups had the potential to utilize large-scale real-world data since collection required specialized and costly hardware. To facilitate experimental research with real data, we created **OpenSky**, a sensor network based on low-cost off-the-shelf equipment connected over the Internet [1]. It currently covers more than 700,000 km^2, seeing around 30% of Europe's commercial air traffic, and makes it possible to analyze ADS-B messages and related metadata.

While similar, commercial, services using ADS-B messages to visualise flight tracks have been available on the Internet, none of them store and provide the valuable raw data required for in-depth research. Therefore, we have made OpenSky an open network that collects and stores all ADS-B traffic as it is being captured. We have been deploying sensor nodes in Central Europe (see Fig. 1b) for a view of the coverage), utilizing volunteers who install sensors at their homes and institutions and deliver their data over the Internet. OpenSky uses cheap off-the-shelf sensors, creating a low barrier of entry for volunteers. During OpenSky's operational phase, we have been working with the data in different ways, including but not limited to:

- **Error and fault diagnosis**: OpenSky can help to discover misbehaving and erroneous transponders which do not comply with standards, detecting safety-related issues prior to wide-scale adoption.
- **Performance evaluation**: OpenSky can assess the performance of the ADS-B channel such as the message loss rate or the number of collisions at various locations and times, identifying bottlenecks in the system capacity.
- **Localization**: When a signal is received by four or more sensors, the position of the sender can be calculated independently through multilateration, providing a backup system to verify the location of an aircraft.
- **Security research**: Several security vulnerabilities have been shown to affect ADS-B which can not be easily addressed as the application of cryptography would require an expensive new system and decades of standardisation. With OpenSky, ground-based attack detection methods and security mitigation techniques can be explored.

The network has been operational for over two years, collecting billions of ADS-B messages for further analysis. All of the stored data is accessible to the volunteers contributing with their sensors, and to anyone else on request on `http://opensky-network.org`.

OpenSky is a joint effort between the University of Oxford, the University of Kaiserslautern, Germany and armasuisse, Switzerland.

References

1. Schäfer, M., Strohmeier, M., Lenders, V., Martinovic, I., and Wilhelm, M. Bringing Up OpenSky: A Large-scale ADS-B Sensor Network for Research. In *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)* (Apr. 2014).
2. Strohmeier, M., Lenders, V., and Martinovic, I. On the Security of the Automatic Dependent Surveillance-Broadcast Protocol. *Accepted for Publication in IEEE Communications Surveys and Tutorials* (2014).
3. Strohmeier, M., Schäfer, M., Lenders, V., and Martinovic, I. Realities and Challenges of NextGen Air Traffic Management: The Case of ADS-B. *Communications Magazine, IEEE 52*, 5 (2014).

Identity Security in Cyberspace - A data-reachability Model

Elizabeth Phillips
Oxford University
Worcester College, Walton Street
Oxford, UK
(+44)7454 817205
elizabeth.phillips@cybersecurity.ox.ac.uk

1. INTRODUCTION

Privacy and security within Online Social Networks (OSNs) has become a major concern over recent years. As individuals continue to actively use and engage with these mediums, we need to appreciate what unknown risks users face as a result of unchecked publishing and sharing of content/ information in this space. There are numerous tools and methods under development that claim to facilitate the extraction of specific classes of personal data from online sources, either directly or through correlation across a range of inputs.

In this poster we present a model which specifically aims to understand the potential risks faced should all of these tools and methods be accessible to a malicious entity. The model enables easy and direct capture of the data extraction methods through the encoding of a data-reachability matrix for which each row represents an inference or data-derivation step. In essence, we view this work as a key method by which we might make cyber risk more tangible to users of OSNs.

For a source element e_a of type a, a transform $t_{a \to b}$ allows us to generate an element e_b of type b. In order to collate data points of an individual together, we define a Characteristic C as a multiset of elements of the same type a and a Superidentity S as the set of characteristics belonging to a particular identity. Figure 1 shows an example of how a superidentity can be established in two rounds of inferences.

2. BACKGROUND AND RELATED WORK

The types of obtainable information is essentially unlimited within the parameters of what an individual chooses to reveal about themselves and their peers. Gross and Acquisti [2], for instance, found this included dates of birth, addresses, phone numbers, relationship status, views and interests, and screen names on other online social network (OSN) sites. A more recent study [3] also highlights these and more attributes (e.g., hobbies, home town, education, favourites, religious views and political direction), and how openly they are shared by users in four popular networks.

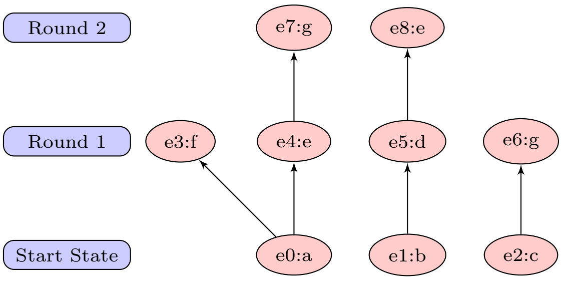


Figure 1: A diagram reflecting an example use of the transforms to go from $S = \{e_0 : a, e_1 : b, e_2 : f\}$ to $S = \{e_0 : a, e_1 : b, e_2 : f, e3 : f, e4 : e, e5 : d, e6 : g, e7 : g, e8 : e\}$ after just 2 rounds of enrichment.

3. PROJECT AIMS

The project aimed to create a holistic model in which we can assess the risk exposure of an individual online by traversing our model to infer new information .

4. APPROACH AND UNIQUENESS

Even though tools such as Spokeo exist to correlate some structured data sources, the approach undertaken is unique in its completeness and the transitive closure of our matrix allows for a comprehensive analysis of the transforms and data available.

4.1 Approach

We set out to tackle the problem of collecting transforms into 3 stages.

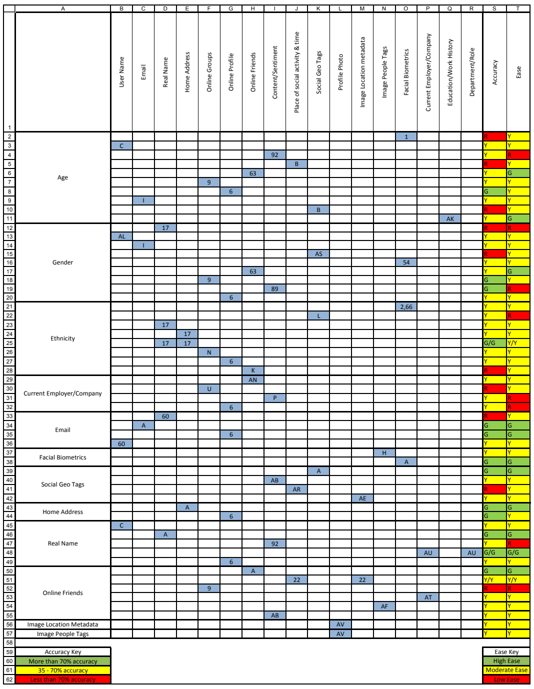
Step 1 :- Define the Data Points

- These are attributes of an individual which can be found online e.g. DOB, friends, location etc.
- These were collected from commonly available information on OSNs and potentially sensitive personal attributes in the offline world.

Step 2 :- Identify transforms that allow reachability

- These are the transforms that allow us to use one or more data points to reach a new data point.
- An extensive literature review of articles to justify these inferences were conducted alongside experiments to predict new inferences.

Step 3 :- Model the reachability of the data points.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		User Name	Email	Real Name	Home Address	Online Groups	Online Profile	Online Friends	Content/Sentiment	Place of social activity & time	Social Geo Tags	Profile Photo	Image Location metadata	Image People Tags	Facial Biometrics	Current Employer/Company	Education/Work History	Department/Role	Accuracy	Ease
2	Age														1				R	Y
3		C																	Y	Y
4									92										Y	R
5										B									R	Y
6								63											Y	G
7						9													Y	Y
8							6												G	Y
9			I																Y	Y
10											B								R	Y
11																	AK		Y	G
12	Gender			17															R	R
13		AL																	Y	Y
14			I																Y	Y
15											AS								R	Y
16															54				Y	Y
17								63											Y	G
18						9													G	Y
19									89										G	R
20							6												Y	Y
21	Ethnicity														2,66				Y	Y
22											L								Y	R
23				17															Y	Y
24					17														Y	Y
25				17	17														G/G	Y/Y
26						N													Y	Y
27							6												Y	Y
28								K											R	Y
29	Current Employer/Company							AN											Y	Y
30						U													R	Y
31									P										Y	R
32							6												Y	R
33	Email			60															R	Y
34			A																G	G
35							6												G	G
36		60																	Y	Y
37	Facial Biometrics													H					Y	Y
38															A				G	G
39	Social Geo Tags										A								G	G
40									AB										Y	Y
41										AR									R	Y
42													AE						Y	Y
43	Home Address				A														G	G
44							6												G	Y
45	Real Name	C																	Y	Y
46				A															G	G
47									92										Y	R
48																AU		AU	G/G	G/G
49							6												Y	Y
50	Online Friends							A											G	G
51										22			22						Y/Y	Y/Y
52						9													R	R
53																AT			Y	Y
54														AF					Y	Y
55									AB										Y	Y
56	Image Location Metadata											AV							Y	Y
57	Image People Tags											AV							Y	Y
58																				
59	Accuracy Key																		Ease Key	
60	More than 70% accuracy																		High Ease	
61	35 - 70% accuracy																		Moderate Ease	
62	Less than 70% accuracy																		Low Ease	

Figure 2: An extract of our final model. In order to infer a data point in the row you need to have every element in the column. Row 25 shows that by using reference 17 you can combine Real Name and Home address to infer ethnicity with high accuracy and moderate ease.

4.2 The Model

The reachability of our data points is modelled in a matrix format for ease of use.

We calculate the transitive closure (every single combination of inference steps) to establish all possible data points which can be gathered from an initial set (see figure 3). Figure 2 illustrates an extract from the matrix created.

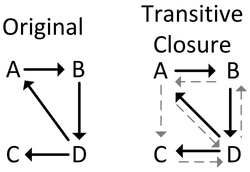


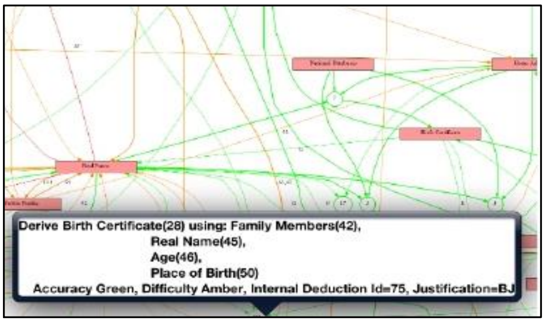
Figure 3: The transitive closure of an original matrix highlighting the 5 newly obtained inferences in grey

4.3 Evaluation

To evaluate our approach to the problem, we designed a working experiment to evaluate the effectiveness of our approach. In order to help guide the users through the inferences available with their current data points, tool support shown in figure 4 was created to allow us to evaluate our inferences and highlight any areas for further research. In order to evaluate the validity of our inferences, we used our tool to help inform us when inferring new information on a real target.

4.4 Benefits

This approach allowed us to ask questions relating to how much information might be gathered in order to calculate the exposure risk of an individual in a systematic way.



Derive Birth Certificate(28) using: Family Members(42),
Real Name(45),
Age(46),
Place of Birth(50)
Accuracy Green, Difficulty Amber, Internal Deduction Id=75, Justification=BJ

Figure 4: A sample black and white graphic (.pdf format).

5. RESULTS AND CONTRIBUTIONS

5.1 Results

Using our tool with our matrix we were able to use our participant's name alone as our initial seed to infer 42 verified personally identifiable information (PII) associated with the individual using only 4 rounds of inferences. The tool allowed us to direct our analysis and allowed us to infer new PII.

Table 1: Table captions should be placed above the table.

Round	No. new data points collected	Total Data Points
Start	1	1
Round 1	7	8
Round 2	13	21
Round 3	16	37
Round 4	5	42

5.2 Contributions

- **Contribution 1.** We were able to highlight the exposure an individual has on their online social networks which allow us to link a professional and social persona.

- **Contribution 2.** By calculating the transitive closure of our matrix and by using the breadth of sources and tools available, we were able to create a comprehensive tool-supported way of inferring PII from multiple sources.

- **Contribution 3.** The ease and accuracy heuristics allow us to tailor the effectiveness of our transforms and allow us to prioritise one source of data over another.

- **Contribution 4.** This added granularity allows users to configure the ease and accuracy to represent their own perspectives, inferences and data points.

- **Contribution 5.** The approach allows us to not only assess the level of vulnerability but also highlight any critical inferences that have the most significant impact.

6. FUTURE WORK

In order to overcome the difficulty of assigning just 1 level of ease for a transform, we will split the ease dimension into 15 sub-dimensions to provide further granularity. We are also comparing different mathematical methods to create a transitive score for chained transforms.

7. REFERENCES

[1] S. Creese, M. Goldsmith, J. R. Nurse, and E. Phillips. A data-reachability model for elucidating privacy and security risks related to the use of online social networks. In *11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-12)*, pages 1124–1131. IEEE, 2012. bibtex: Creese2012.

[2] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM.

[3] S. Labitzke, I. Taranu, and H. Hartenstein. What your friends tell others about you: Low cost linkability of social network profiles. In *Proc. 5th International ACM Workshop on Social Network Mining and Analysis, San Diego, CA, USA*, 2011.

Deniability in Security Protocols

Research Proposal

Katriel Cohn-Gordon
Supervised by Dr Cas Cremers

Cyber Security CDT
Robert Hooke Building, Parks Road, Oxford

To build secure protocols we put together a number of building blocks, of which key exchange is one of the most used and useful: two parties exchange some messages, after which they and only they share a secret key. We can judge these key exchange protocols by various criteria: the secrecy of the resultant key, the authenticity of the participants, and the number of messages or mathematical operations needed, among others. For example, the SIGMA protocol of Krawczyk [3] (Figure 1, standardised as part of the Internet Key Exchange RFC), uses signatures together with a keyed hash function in order to authenticate both parties to each other, as well as to ensure secrecy of the key it generates.

Another criterion that can be used to judge key exchange protocols is how much of a trace they leave on the network: whether they leak the intended party to a communication, for instance, or provide proof that one of the parties was alive and transmitting at a given time. The 21st-century Zeitgeist provides a wealth of examples: a whistleblower in a large organisation probably does not wish to prove to the world that she is communicating with a journalist, nor a politician that she is leaking data to a news agency, nor a citizen of a repressive regime that she is accessing a banned website.

In practice, these traces often consist of cryptographic signatures of interim messages, which serve as an extremely useful tool for authentication but by construction can be verified by anyone with access to public keys. Let us look in more detail at SIGMA, whose formal definition is detailed in Figure 1.

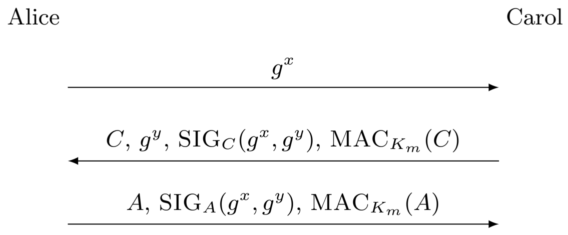

Alice
Carol
g^x
C, g^y, $\mathrm{SIG}_C(g^x, g^y)$, $\mathrm{MAC}_{K_m}(C)$
A, $\mathrm{SIG}_A(g^x, g^y)$, $\mathrm{MAC}_{K_m}(A)$

Fig. 1. SIGMA

Here g is an element of some fixed group, x and y are the random secret keys chosen by Alice and Carol respectively, SIG is a cryptographic signature with the long-term secret key of the respective party, and MAC is a “message authentication code” or keyed hash function with message key K_m derived from g^{xy}. The security stems from the “computational Diffie-Hellman hypothesis” that computing g^{xy} from g, g^x and g^y is difficult unless you know x or y, in which case it is easy.

In the third message of SIGMA, Alice broadcasts $\mathrm{SIG}_A(g^x, g^y)$. Anyone can verify that this is a valid signature which only Alice could have produced, and therefore that she must at some point have participated in a SIGMA key exchange. Moreover, if Carol chooses g^y to encode e.g. the hash of today’s *New York Times*, then the signature proves that Alice must have been alive and signing messages today. It does not, however, prove that Alice intended to exchange a key *with Carol*: she would have created the same signature to talk to anyone.

It seems, then, that there is a sequence of potential types of deniability. In increasing order of strength, the traces left on the network ("the protocol transcript") could prove that Alice deliberately exchanged a key with Carol today, that she exchanged a key with someone today, that she exchanged a key with someone at some point, or perhaps not even that. SIGMA appears to satisfy the penultimate of these.

Some notions of deniability are closely related to those of signing adversarial data; that is, whether a protocol role instance is willing to provide a signature on some piece of data provided by a potential adversary. We propose to formalise this link, together with some related definitions, in order to pin down precisely which properties of protocols provide or prevent deniability. This will build on the work of Cremers and Feltz [2], proposing new definitions and proofs applied to current and proposed key exchange protocols.

More generally, there are plenty of different definitions of deniability in the literature, depending on the threat model, the abilities of the judge, the particular level of deniability required and more. For instance, if Carol is assumed to be honest then it is easier for Alice to achieve a notion of deniability than if she is assumed to be malicious; likewise, if the judge (whom Carol is trying to convince of the authenticity of the transcript) is allowed actively to interfere in the original protocol run then deniability is harder to achieve.

In this project, we aim to state various of these definitions in a comparable fashion, and show either incomparability or implication between them. Ideally, these will also tease out the tradeoffs between the various key exchange criteria: "it is impossible to achieve X level of deniability and still maintain authenticity", or "any protocol that achieves Y must be subject to the following genre of attack". Making these tradeoffs explicit, and thus pinpointing the boundaries of the protocol space, should help not only with choosing which to use for particular applications, but also with the development of novel protocols that reach them.

References

1. Boyd, C., Mao, W., and Paterson, K. G. Deniable authenticated key establishment for internet protocols. In *11th International Workshop on Security Protocols* (2003), Springer, p. 255271.
2. Cremers, C., and Feltz, M. One-round strongly secure key exchange with perfect forward secrecy and deniability. Tech. Rep. 300, 2011.
3. Krawczyk, H. SIGMA: the SIGn-and-MAc approach to authenticated diffie-hellman and its use in the IKE protocols. In *Advances in Cryptology - CRYPTO 2003*, D. Boneh, Ed., no. 2729 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Jan. 2003, pp. 400–425.
4. Raimondo, M. D., Gennaro, R., and Krawczyk, H. Deniable authentication and key exchange. Tech. Rep. 280, 2006.

Effective Verification of Low-Level Software with Nested Interrupts

Lihao Liang
Department of Computer Science, University of Oxford

The interrupt mechanism enables timely response to outside stimuli in a power-efficient way, therefore is a key design primitive for embedded software that interacts closely with hardware. Interrupts are widely used in all styles of computing platforms, including safety-critical embedded software, low-power mobile platforms, and high-end information systems.

But interrupt-driven programs are difficult to engineer. Device drivers, typical examples of software that uses interrupts heavily, contain most of the faults in the Linux Kernel [8]. Interrupts can arrive at arbitrary times, leading to an explosion in the number of cases to be considered in validation. Most existing approaches to validating interrupt-driven code rely on testing. In the case of nested interrupts, testing is particularly ineffective, as interspersions of interrupts within a run of the code grows exponentially in the number of interrupts that occur. Bugs are therefore easily missed, and any errors that are observed are difficult to reproduce. This motivates a formal approach to validating software with nested interrupts.

Our Technique

We consider prioritised, preemptive scheduling policies as for instance, provided by the x86 architecture, and most microcontrollers architectures, e.g. 8051. Interrupt priorities can be either fixed (e.g. AVR1305) or configurable, often by assigning interrupt levels to individual interrupts (e.g. ARM Cortex-M, AT89Cxx). In all these architectures, the handling of lower-priority interrupts, when a higher-priority interrupt arrives, will be suspended until the handling of the higher-priority one is completed.

Our contribution is a novel method to verify programs with nested interrupts in the form of symbolic execution based on a partial-order encoding that precisely models the interleaving semantics of nested interrupts with priorities. The idea is to translate a program into atomic memory read/write events, and then encode all interleavings of these events that can possibly occur as a symbolic partial-order in a SAT/SMT formula. A satisfying assignment to the variables in the formula corresponds to an error trace in the interrupt-driven program. Figure 1 gives an overview of our technique.

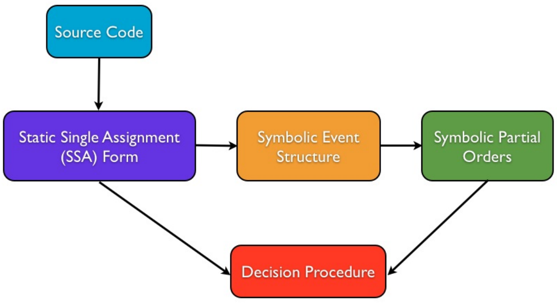


Fig. 1. Overview of Our Technique

The encoding is a logical formula of the form `ssa` ∧ `pord`, where `ssa` encodes the data and control structure of the interrupt-driven program under analysis, and `pord` encodes all possible interleavings of the different interrupt handlers in the program. The first conjunct `ssa` is built in the same way as [2], which is based on a variant of *static single assignment* (SSA) form as described in [4]. Once the SSA encoding `ssa` is obtained, we construct a collection of *symbolic events* based on the occurrences of variables being read or written. By referring to symbolic events, we can construct in the second conjunct `pord` a symbolic partial-order that captures precisely the subset of interleavings of reads and writes of the original program required to model nested interrupts faithfully.

Experimental Studies

To evaluate the performance of our partial-order encoding for nested interrupts, we implemented a prototype tool i-CBMC, an extension of CBMC [3]. We experimentally compare our new technique with conventional approaches based on source-to-source transformation to sequential [5] or multi-threaded programs [9] that can be handled by off-the-shelf verification tools such as BLAST [10], UFO [1], Impara [11], ESBMC [7], and CBMC [3].

We assess the effectiveness of each method, on the following benchmarks derived from real-world embedded systems code and Linux device drivers: Logger (172 lines of code), parts of the firmware of a temperature logging device from a major industrial enterprise; Blink (2652 LOC), an LED application in the TinyOS distribution; Brake (2473 LOC), a Simulink model of a brake-by-wire system of Volvo Technology AB; RcCore (7035 LOC), a Linux device driver for a remote control together with a model of the Linux Kernel.

The experimental results demonstrate that our partial-order encoding is the most efficient and effective method to verify software with nested interrupts, outperforming the competing approaches significantly [6]. Our novel encoding provides the first demonstrated method for effective formal verification of low-level embedded software with nested interrupts.

References

1. Albarghouthi, A., and Gurfinkel, A. UFO: Abstract interpretation and interpolants - (competition contribution). In *TACAS* (2014).
2. Alglave, J., Kroening, D., and Tautschnig, M. Partial orders for efficient Bounded Model Checking of concurrent software. In *CAV* (2013), vol. 8044 of *LNCS*, pp. 141–157.
3. Clarke, E., Kroening, D., and Lerda, F. A tool for checking ANSI-C programs. In *TACAS* (2004), pp. 168–176.
4. Clarke, E., Kroening, D., and Yorav, K. Behavioral consistency of C and Verilog programs using bounded model checking. In *DAC* (2003), pp. 368–371.
5. Kidd, N., Jagannathan, S., and Vitek, J. One stack to run them all – reducing concurrent analysis to sequential analysis under priority scheduling. In *SPIN* (2010), pp. 245–261.
6. Kroening, D., Liang, L., Melham, T., Schrammel, P., and Tautschnig, M. Effective verification of low-level software with nested interrupts. In *submission.*
7. Morse, J., Ramalho, M., Cordeiro, L., Nicole, D., and Fischer, B. ESBMC 1.22 - (competition contribution). In *TACAS* (2014).
8. Palix, N., Thomas, G., Saha, S., Calvès, C., Lawall, J., and Muller, G. Faults in Linux: Ten years later. In *ASPLOS* (2011), pp. 305–318.
9. Regehr, J., and Cooprider, N. Interrupt verification via thread verification. *ENTCS 174*, 9 (2007), 139–150.
10. Shved, P., Mandrykin, M., and Mutilin, V. Predicate analysis with BLAST 2.7.2 - (competition contribution). In *TACAS* (2014).
11. Wachter, B., Kroening, D., and Ouaknine, J. Verifying multi-threaded software with Impact. In *FMCAD* (2013), pp. 210–217.

Search as a Group: Query Answering for the Semantic Web under Group Preferences

Oana Tifrea-Marciuska

Department of Computer Science, University of Oxford
oana.tifrea@cs.ox.ac.uk

In the recent years, several important changes have been taking place on the classical Web. First, the so-called Web of Data is increasingly being realized as a special case of the Semantic Web. Second, as a part of the Social Web, users are acting more and more as first-class citizens in the creation and delivery of contents on the Web. The combination of these two technological waves is called the *Social Semantic Web* (or also *Web 3.0*), where the classical Web of interlinked documents is more and more turning into (i) semantic data and tags constrained by ontologies, and (ii) social data, such as connections, interactions, reviews, and tags. The Web is thus shifting away from data on linked Web pages towards fewer such interlinked data in social networks on the Web relative to underlying ontologies. This requires new technologies for search and query answering, where the ranking of search results is not solely based on the link structure between Web pages anymore, but on the information available in the Social Semantic Web — in particular, the underlying ontological knowledge present in user-created content, as well as the user's preferences implicitly or explicitly present in such content.

Modeling the preferences of a group of users is also an important research topic in its own right. With the growth of social media, people post their preferences and expect to get personalized information. Moreover, people use social networks as a tool to organize events, where it is required to combine the individual preferences and suggest items obtained from aggregated user preferences. For example, if there is a movie night of friends, family trip, or dinner with working colleagues, one has to decide which is the ideal movie or location for the group, given the preferences of each member. To address this problem, individual user preferences can be adopted and then aggregated to group preferences. However, this comes along with two additional challenges. The first challenge is to define a group preference semantics that solves the possible *disagreement* among users (a system should return results in such a way that each individual benefits from the result). For example, people (even friends) often have different tastes in restaurants. The second challenge is to allow for efficient algorithms, e.g., to compute efficiently the answers to queries under aggregated group preferences.

Another aspect that has become increasingly important in recent times is that of uncertainty management, since uncertainty can arise due to many uncontrollable factors. For example when we would need to find out what is the probability that a certain restaurant is a good location for today's dinner we could use the social content (review, ratings) from different websites, that contains uncertain information.

We introduce GPP-Datalog+/−, a language that combines the Datalog+/− ontology language (which is more expressive than DL-Lite and has a more compact representation of the attributes of concepts and roles) with group preferences and probabilistic uncertainty. To our knowledge, this is the first combination of ontology languages with group preferences. The preference and the probabilistic models are assumed to represent the preferences of a group of users and the uncertainty in the domain, respectively.

We formalize the notion of k-rank query answering based on operators for merging single-user and probability-based preferences (in the form of a strict partial and a weak order, respectively), and aggregating multiple single-user preference relations. We analyze two approaches to computing an answer to a k-rank query that are suitable for partially ordered sets of preferences: collapse to single user (CSU), which constructs a single virtual user that aggregates the preferences of all the individuals from the group and the k-rank answers are computed over this new preference relation; and voting-based aggregation, where k-rankings are computed first for each individual user and then aggregation techniques based on voting strategies are used to aggregate the answers and obtain a single k-ranking.

Based on an algorithm for the above preference merging and aggregation, we give algorithms for answering k-rank queries for DAQs (disjunctions of atomic queries), which generalize top-k queries based on the iterative computation of classical skyline answers. We show that answering DAQs in GPP-Datalog+/− is possible in polynomial time in the data complexity modulo the cost of computing probabilities.

Finally, we developed a prototype implementation of a group preference-based query answering system, and conducted a series of experiments based on real-world ontological data and preference models derived from information gathered from real users. The results (on the running time of our algorithms and the quality of their results) show in particular that the strategies proposed and developed in this work are computationally feasible and semantically reasonable in practice. Moreover, we compared the strategies we presented in terms of efficiency and the quality of the results. From our quality evaluation, it seems that plurality is the most similar to what individual users want rather than collapse to single user.

Topics for future work include further exploration of the similarity of preference aggregation and merging strategies to human judgment; this will shed light on how well-suited each of them is as a general aggregation strategy for search and query answering in the Social Semantic Web. Related to this effort, our experimental evaluation shows that new methods for measuring user's satisfaction within a group should be developed, perhaps based on the difference between least and most satisfied users, and incorporated into our framework. Another interesting vein for future development is deriving explanations along with the answers to queries in social situations, it is likely that users satisfaction with the answer is tied to how it was produced; for instance, if close friends heavily influenced the result, the user will probably respond better than if the result was influenced by strangers. Another topic of future research is to more deeply explore the computational aspects of query answering in our approach and to explore whether it can be extended to more general queries.

Parts of this work can be found in:

- Lukasiewicz, T., Martinez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Ontology-based query answering with group preferences. Technical report CS-14-02. (2014).
- Lukasiewicz, T., Martinez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Group preferences for query answering in Datalog+/− ontologies. In:Proceedings of the 7th International Conference on Scalable Uncertainty Management SUM 2013 Washington DC USA September 16-18 2013. Vol. 8078 of Lecture Notes in Computer Science. Pages 360-373. Springer. (2013).
- Lukasiewicz, T., Martinez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Query answering in Datalog+/− ontologies under group preferences and probabilistic uncertainty. In: Proceedings of the 2nd International Workshop on Data Management in the Social Semantic Web DMSSW 2013 Aalborg Denmark July 8 2013. Vol. 8295 of Lecture Notes in Computer Science. Pages 192-206. Springer. (2013).
- Lukasiewicz, T., Martinez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Query answering in probabilistic Datalog+/− ontologies under group preferences. In: Proceedings of the 2013 IEEE/WIC/ACM International Conferences on Web Intelligence WI 2013 Atlanta GA USA November 1720 2013. Pages 171-178. IEEE Computer Society. (2013).

Web Object Matching

An analysis of duplicates from web extracted records

Stefano Ortona

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

Today the web has become the largest available source of information and a primary source of structured data. Unfortunately, the vast majority of such data is designed for human fruition only. Data is published via HTML markup and visual styling, without the possibility to access via APIs and query endpoints as it is done by large aggregators such as Amazon or Tripadvisor. The automatic extraction of structured data is a problem of central interest and has been widely investigated [1, 2, 4]. However, after the extraction, the problem of identifying duplicates among the extracted web records must be solved in order to produce clean data for users and applications. Even though this problem, known as record linkage or record matching, has been the focus of several studies in the database community (see [3] for a survey), only few works have tried to address the same problem in the web context [5, 7]. This work investigates the problem of *web object matching*, i.e., identifying duplicates among records extracted from the web. This problem is a small step of the wider problem known as data integration [6], where the aim is not only to identify duplicates, but also to aggregate and integrate data coming from different sources in order to produce a unique, complete and accurate knowledge of the records of interest. In this work we focus primarily on the deduplication, leaving the integration problem as a future work.

Record matching is a well known problem for databases integration. Records to be matched are published from different sources where each source has its own publishing model (*local schema*). Therefore some attributes might be omitted by a source because considered irrelevant, same attributes might be called with different names by two different sources and we might extract different values from different sources for the same attribute. All these issues represent the normal setting for a classic record matching problem in the database scenario. We will show that on the web, the problem of identifying duplicates is harder than what we can expect in a normal database. The target of the integration are records coming from extraction systems that, rather than conventional databases or APIs, might introduce semantic errors that are not due to a problem in the source. Most of the previous approaches ([5, 7]) rely on the fact that records to be matched contain the correct information and misalignments are mostly due to timeliness, denormalisation representational discrepancies, and synthetic errors due to copying rules and data entry. In other words, the information is in some form correct and most importantly, it appears with the "schema". In web object matching, records often require a schema-level reconciliation before the actual deduplication, in order to validate (and possibly correct) the values of the extracted records.

This work presents an approach that performs a repair step before the actual deduplication. We aim at a totally unsupervised approach that involves a component which is an expert on the application domain, the *oracle*. Our oracle is an entity extraction system that is able to identify semantic annotations on the extracted records and use those annotations to adjust the extracted values. The oracle is able to identify the relevant entities (attributes) on the extracted values. After the entities have been identified the records can be repaired by aligning these entities with the original schema of the relation. We present a framework that works not only on a per record fashion (as most of other approaches), but also on per

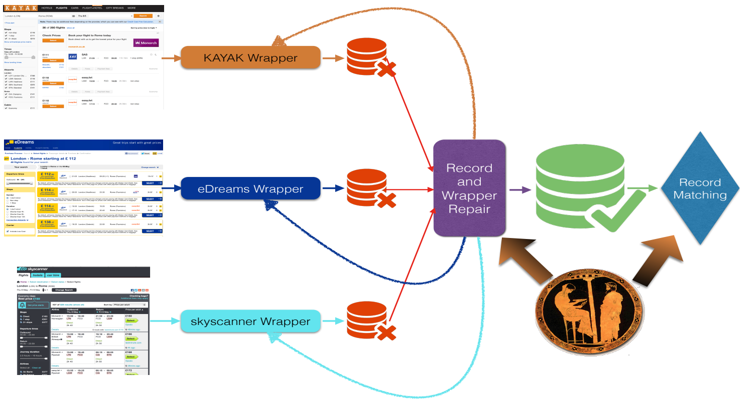


Fig. 1. System architecture overview

source basis which allows detecting and possibly correcting systematic errors for an entire source, that means repairing the extraction system directly rather than individual records. As far as we know this is the first work that involves a repair strategy (both at record and source level) before the actual matching.

Figure 1 shows the architecture of the system. The input of the system is a set of records (along with the extraction rules), extracted from multiple sources of the same domain, that are potentially broken. In these broken records we might expect things like null values, values that contain noise other than the interesting information, values that have been extracted in the wrong attribute position. The first component of our system is the repair step. The goal of this step is to re-adjust the broken records and at the same time to repair the extraction rules (anytime that an error is systematic, that means the same error is repeated for all the records in the source). After the records have been repaired, the system pipeline proceeds with the deduplication step, which has been facilitated by the previous component. Both of these steps involve the use of the oracle (represented as oracle of Delphi in the figure), as source of knowledge of the domain of interest (flights in the figure example).

References

1. Crescenzi, V., Merialdo, P., and Qiu, D. A framework for learning web wrappers from the crowd. In *Proceedings of WWW* (2013).
2. Dalvi, N., Kumar, R., and Soliman, M. Automatic wrappers for large scale web extraction. *Proceedings of the VLDB* (2011).
3. Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. Duplicate record detection: A survey. *TKDE* (2007).
4. Furche, T., Gottlob, G., Grasso, G., Gunes, O., Guo, X., Kravchenko, A., Orsi, G., Schallhart, C., Sellers, A., and Wang, C. Diadem: domain-centric, intelligent, automated data extraction methodology. In *Proceedings of WWW* (2012).
5. Kannan, A., Givoni, I. E., Agrawal, R., and Fuxman, A. Matching unstructured product offers to structured product specifications. In *Proceedings of SIGKDD* (2011).
6. Lenzerini, M. Data integration: A theoretical perspective. In *Proceedings of SIGMOD* (2002).
7. Su, W., Wang, J., and Lochovsky, F. H. Record matching over query results from multiple web databases. *TKDE* (2010).

Generic Inference and Machine Learning

Distributed inference techniques applied to machine learning

Abhishek Dasgupta

Oxford University Department of Computer Science
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

In recent years, we have seen the rise of multi-core processors and an increasing emphasis on parallel computation as the sequential speed of single processors plateau. The existence of large datasets has made it possible to perform statistical inference on them. Until recently, machine learning algorithms were usually developed for sequential architectures, and was dominated by imperative languages like MATLAB; the lack of parallel programming frameworks for machine learning did not pose a problem because of the size of the datasets involved.

The introduction of MapReduce has seen an increasing interest in the development of parallel frameworks targeted towards solving machine learning problems. One of the problems of MapReduce is that it does not work well for problems which have anything but the most trivial dependencies between their data. An example of such a problem is calculating the PageRank of a page, while it is possible to adapt it to MapReduce, a message-passing semantics is in many cases a simpler approach.

In the past couple of years, new frameworks [6, 1, 4, 7] have been developed which try to address the issue. Most of these are based on representing the data in machine learning problems in the form of a graph which represents the dependencies among the data, and using message passing as a core part of the framework, borrowing ideas from previous parallel frameworks based on graphs like Dryad [5] which represents computational dependencies in the form of directed acyclic graphs.

From a theoretical perspective, such frameworks develop theories of parallel programming as applied to machine learning. Most of the frameworks have focused on the practical aspect of language-building.

We focus on inference algorithms, as they are an important subclass of general machine learning algorithms. Frameworks like GraphLab [6] can be used for inference, but their scope is more generic than the inference problem and can be used to structure problems like PageRank as they have been developed as frameworks which structure parallel computation over graphs. In comparision, there are older frameworks like the Bayesian logic Inference Engine (BLOG) and Bayesian inference using Gibbs sampling (BUGS) which do not include parallelism.

In order to develop the unified approach to developing parallel frameworks for the inference problem, we make use of the generic inference framework developed by Kohlas and Pouly [9]. The generic inference framework can describe not only the inference problem for undirected (Markov networks) and directed (Bayesian networks), but also other problems with a similar structure like querying in relational databases. The theory is based on valuation algebras which express the inference problem as a series of atomic combination and projection operations (equivalent to marginalisation operation) of information. In the case of relational databases, this corresponds to the *join* and *select* operations respectively. The benefit of using a generic framework which also utilises message passing like the above mentioned frameworks is that any distributed algorithm for solving the generic inference problem immediately yields equivalent algorithms in all these areas.

Though we use generic inference, our main aim is exploring parallelisation in the inference problem in the domain of machine learning; the other areas under the purview of generic inference shall benefit, but we shall explore those, if time permits, and then only in passing.

We're presently building a generic inference library based on Cloud Haskell [2]. Using a functional programming language offers benefits such as clearer separation of implementation details and the algorithmic aspects. Cloud Haskell is a recently developed framework which ports the Erlang semantics of distributed programming via message passing to Haskell. In contrast to existing software implementations, we shall implement an approximate inference framework as in [3]. The framework will be based on the Bulk Synchronous Parallel theory [10] which has also been the basis of Google's Pregel [7] graph processing framework. Using BSP provides a theoretical foundation and helps analyse the communication costs in the distributed framework, like in [8] which did this for MapReduce. The implementation is in an early stage, and we hope to have a proof-of-concept implementation developed by August 2014.

Why is this project relevant to the machine learning community at large? While most of the contribution of this work is in the development of a robust and extensible functional programming framework which we hope will be a basis for further work in the area of functional programming as applied to machine learning, there is scope for unifying existing algorithms like Gibbs sampling and variational inference under the umbrella of such a generic framework, as well as extending the framework to anytime algorithms which can incrementally compute better approximations to the exact solution. Any progress towards such a goal would be beneficial as efforts can be focused on improving and optimising the generic framework as opposed to specific algorithms.

References

1. The GraphLab abstraction. http://graphlab.org/abstractiononly.pdf, retrieved 2012-06-09.
2. Epstein, J., Black, A. P., and Peyton-Jones, S. Towards Haskell in the cloud. In *Proceedings of the 4th ACM symposium on Haskell* (New York, NY, USA, 2011), Haskell '11, ACM, pp. 118–129.
3. Haenni, R. Ordered valuation algebras: a generic framework for approximating inference. *International journal of approximate reasoning 37*, 1 (2004), 1–41.
4. Haller, P., and Miller, H. Parallelizing machine learning– functionally: A framework and abstractions for parallel graph processing. In *Scala Workshop 2011* (2011).
5. Isard, M., Budiu, M., Yu, Y., Birrell, A., and Fetterly, D. Dryad: distributed data-parallel programs from sequential building blocks. In *Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007* (New York, NY, USA, 2007), EuroSys '07, ACM, pp. 59–72.
6. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., and Hellerstein, J. M. Graphlab: A new parallel framework for machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)* (Catalina Island, California, July 2010).
7. Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., and Czajkowski, G. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 international conference on Management of data* (New York, NY, USA, 2010), SIGMOD '10, ACM, pp. 135–146.
8. Pace, M. F. {BSP} vs mapreduce. *Procedia Computer Science 9*, 0 (2012), 246 – 255. Proceedings of the International Conference on Computational Science, {ICCS} 2012.
9. Pouly, M., and Kohlas, J. *Generic Inference: A Unifying Theory for Automated Reasoning.* Wiley-Blackwell, May 2011.
10. Valiant, L. G. Handbook of theoretical computer science (vol. a). MIT Press, Cambridge, MA, USA, 1990, ch. General purpose parallel architectures, pp. 943–973.

Completeness results for the ZX-calculus for quantum computing

Miriam Backens

Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

The study of quantum computing is the study of computation which makes explicit use of quantum-mechanical phenomena such as entanglement and superposition of states. Where a classical computer performs boolean operations on bits, which take the values 0 or 1, a quantum computer performs unitary operations on qubits, whose states are usually modelled as vectors in a two-dimensional complex Hilbert space, with the operations given by matrices. This approach to the study of quantum mechanics (QM) is called *matrix mechanics*.

A different paradigm for the study of quantum computing is offered by categorical quantum mechanics [1]. Using category theory to investigate quantum systems puts the focus on the transformations rather than the states of the systems. It also allows the introduction of graphical calculi, i.e. diagrammatic languages for describing quantum operations. These graphical calculi come with built-in rewrite rules which allow diagrams to be simplified, or can be used to derive equalities between diagrams.

One such graphical calculus is the ZX-calculus, introduced in [4]. A ZX-calculus diagram consists of nodes connected by edges (see Fig. 1 for examples). Each edge represents a qubit and each node represents a transformation acting on some number of qubits. Diagrams are read from bottom to top; edges going into a node are called inputs and edges going out of a node are called outputs. There are three types of nodes: green nodes with any number of inputs and outputs (cf. Fig. 1a), red nodes with any number of inputs and outputs (Fig. 1b), and Hadamard nodes, which always have one input and one output (Fig. 1c). Green and red nodes may also have an angle attached to them, called the *phase*, a missing phase label indicates an angle of 0. Dangling edges are inputs or outputs for the diagram as a whole. The green and red nodes represent two classes of maps, which are associated with two different bases for the underlying Hilbert space. The Hadamard nodes map from one of these bases to the other.

There are about a dozen basic rewrite rules for the ZX-calculus [5, 6], including the green spider rule which states that two green nodes connected by at least one edge merge, their phases adding. Another example is the colour change rule: a green node with Hadamard nodes on each of its inputs and outputs can be replaced by a red node of the same phase. Fig. 1d shows an example of diagram rewrites using these two rules.

For the ZX-calculus to be a valid alternative way of doing calculations about quantum systems, it needs to have certain properties with respect to matrix mechanics. Firstly, it

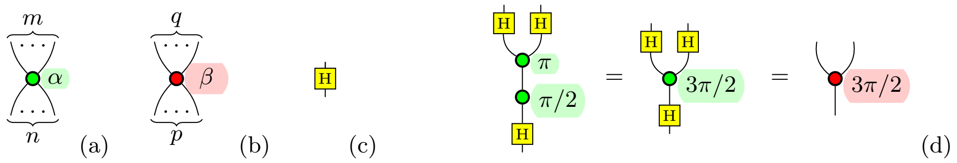


Fig. 1. Examples of ZX-calculus diagrams and equalities: (a) a green node with n inputs, m outputs and phase α; (b) a red node with p inputs, q outputs, and phase β; (c) a Hadamard node; (d) rewriting a ZX-calculus diagram using first the spider rule and then the colour change rule.

should be *universal*, meaning any state of or operation on a quantum system should be expressible as a ZX-calculus diagram. Secondly, any equality between diagrams that can be derived using the rewrite rules should hold true when translated into matrices. This property is *soundness*. Thirdly, whenever two diagram correspond to operators that are equal in matrix mechanics, it should be possible to show that the diagrams are equal using the rewrite rules. This is *completeness*. The ZX-calculus is universal and sound by construction [5]. The question of completeness was unresolved until recently, but it has since been shown that the ZX-calculus is incomplete overall [7]. This result holds even when the diagrams are restricted to be line graphs, i.e. all operations act on a single qubit.

Given this incompleteness proof, there are two ways of getting completeness results anyway: one can either add more rewrite rules until the calculus is complete, or one can restrict the allowed elements so that the calculus only describes a fragment of qubit quantum mechanics. Here, we focus on the latter method. The incompleteness argument relies on the fact that phases can take arbitrary values, thus the easiest way of restricting the ZX-calculus is by restricting the phases. For example, rather than arbitrary angles, they could be constrained to be integer multiples of $\pi/2$. This restricted version of the ZX-calculus does in fact describe a widely studied fragment of QM called *stabilizer quantum mechanics*.

Indeed, we showed that not only does the incompleteness argument fail for the stabilizer ZX-calculus, but the stabilizer ZX-calculus is complete. This is because any stabilizer diagram can be rewritten into a normal form, which is not unique but nevertheless allows straightforward equality-testing [2].

Another complete fragment of the ZX-calculus is the one where all diagrams are line graphs, i.e. every node has exactly one input and one output, and phases are integer multiples of $\pi/4$. For these diagrams there exists a unique normal form, and an efficient procedure for rewriting diagrams to normal form. The set of operators that can be represented as diagrams of this type contains arbitrarily good approximations to any single-qubit unitary operator. We thus say that the ZX-calculus is *approximately complete* for single qubits [3].

These two known-to-be-complete fragments of the ZX-calculus overlap only partially: the former is more restrictive on phases, whereas the latter is more restrictive on diagram structure. An obvious next question is whether the combination of the two—i.e. arbitrary graphs with phases that are integer multiples of $\pi/4$—is complete. Research into this topic is ongoing.

References

1. Abramsky, S., and Coecke, B. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science (LICS'04)* (July 2004), pp. 415 – 425.
2. Backens, M. The ZX-calculus is complete for stabilizer quantum mechanics. arXiv e-print 1307.7025, July 2013.
3. Backens, M. The ZX-calculus is approximately complete for single qubits. In *QPL 2014* (June 2014). To appear.
4. Coecke, B., and Duncan, R. Interacting quantum observables. In *Automata, Languages and Programming*, vol. 5126. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 298–310.
5. Coecke, B., and Duncan, R. Interacting quantum observables: categorical algebra and diagrammatics. *New Journal of Physics 13*, 4 (Apr. 2011), 043016.
6. Duncan, R., and Perdrix, S. Graph states and the necessity of Euler decomposition. In *Mathematical Theory and Computational Practice*, vol. 5635. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 167–177.
7. Schröder, C., and Zamdzhiev, V. Private communication. Oct. 2013.