Computing and Information Systems

3RD ANNUAL DOCTORAL COLLOQUIUM

Program and Proceedings

22nd July 2015
Welcome to the CIS 2015 Doctoral Colloquium!

For the third consecutive year, The Department of Computing & Information Systems (CIS) at The University of Melbourne, is hosting the Doctoral Colloquium (DC) and on behalf of the 3rd CIS DC organizing committee, I am honoured and delighted to welcome each of you to this annual event.

The CIS DC aims to create opportunities for graduate research students by motivating them to explore new research areas and foster their awareness of recent advances and updates offered by other students. The colloquium targets a broad range of participants, from new candidates to those who are ready to submit their thesis. By providing a collaborative and interactive environment bringing together academics, industry partners and students, all participants are able to benefit from their involvement in this event. Paper presentations, poster sessions and flash talks are included in the program to augment the essential social and presentation skills of our graduate research students.

We are excited to report a 32% increase in the total number of paper and poster submissions this year, which is more than double the number of submissions received for the inaugural colloquium held in 2013.

The success of the CIS DC 2015 is due to the hard work of staff and graduate research students in CIS who have worked with us in planning and organizing both the technical program and supporting social arrangements. In particular, I would like to thank Justin Zobel, Jan Schroeder, Rhonda Smithies and Julie Ireland for their much valued advice and support. Needless to say, this colloquium could not happen without the involvement of our graduate research students who have volunteered their time by serving on the organising committee and we would like to acknowledge their hard work, energy and passion for this event. Finally, we would like to thank our sponsors, Google, Microsoft, the Microsoft Research Centre for Social Natural User Interfaces, the Melbourne School of Information and the Melbourne School of Engineering, for their support of the colloquium.

We hope that you will find attending this year’s CIS DC a rewarding experience.

Farzad Khodadadi
Committee Chair
Conference Location: Alan Gilbert Building

Building Number: 104
Street Address: Grattan Street
Campus Map reference: M13

Image source: en.wikimedia.org

Map of Conference Venue

Leaflet | Data, imagery and map information provided by MapQuest, OpenStreetMap and contributors.
### Conference Program

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity 1</th>
<th>Activity 2</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30-9:00</td>
<td>Registration</td>
<td></td>
<td>Foyer, Level 1 (Adjacent to Lecture Theatre 3)</td>
</tr>
<tr>
<td>9:00-9:15</td>
<td>Opening and Welcome Address Prof. Tim Baldwin</td>
<td></td>
<td>Lecture Theatre 1 (Ground Floor)</td>
</tr>
<tr>
<td>9:15-10:15</td>
<td>Paper Session 1A (Machine Learning)</td>
<td>Paper Session 1B (Human Computer Interaction and Design)</td>
<td>Lecture Theatre 1 &amp; 3</td>
</tr>
<tr>
<td>10:15-10:45</td>
<td>Morning Tea</td>
<td></td>
<td>Foyer, Level 1</td>
</tr>
<tr>
<td>10:45-11:30</td>
<td>Keynote: System U to Watson Personality Insights: The Transformation of A Research Project into a Commercial Offering Dr Anna Phan</td>
<td></td>
<td>Lecture Theatre 3</td>
</tr>
<tr>
<td>11:30-12:30</td>
<td>Paper Session 2A (Machine Learning)</td>
<td>Paper Session 2B (Business Information)</td>
<td>Lecture Theatre 1 &amp; 3</td>
</tr>
<tr>
<td>12:30-13:30</td>
<td>Mid Day Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:30-14:30</td>
<td>Paper Session 3A (Computational Bioinformatics and Health Information Systems)</td>
<td>Paper Session 3B (Distributed and Cloud Computing)</td>
<td>Lecture Theatre 1 &amp; 3</td>
</tr>
<tr>
<td>14:30-15:00</td>
<td>Afternoon Tea</td>
<td></td>
<td>Foyer, Level 1</td>
</tr>
<tr>
<td>15:00-16:00</td>
<td>Paper Session 4A (Natural Language Processing, Databases and Business Information)</td>
<td>Paper Session 4B (Optimization and Programming Languages)</td>
<td>Lecture Theatre 1 &amp; 3</td>
</tr>
<tr>
<td>16:00-17:30</td>
<td>Flash Talks</td>
<td>Poster Presentations</td>
<td>Lecture Theatre 3 and Foyer, Level 1</td>
</tr>
<tr>
<td>16:45-18:30</td>
<td>Refreshments</td>
<td></td>
<td>Foyer, Level 1</td>
</tr>
<tr>
<td>17:30-18:30</td>
<td>Closing and Award Ceremony</td>
<td></td>
<td>Foyer, Level 1</td>
</tr>
</tbody>
</table>
## Session Papers

### Session 1A (Machine Learning)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuo Wang and Richard Sinnott</td>
<td>Privacy-Preserving Location Mining of Tweets</td>
</tr>
<tr>
<td>Elham Naghi Zadeh, James Bailey, Lars Kulik and Egemen Tanin</td>
<td>How Private Can I Be Among Public Users?</td>
</tr>
<tr>
<td>Tuan Doan, Sutharshan Rajasegarar and Christopher Leckie</td>
<td>Modelling Pedestrian Activities in a Dynamic Environment</td>
</tr>
<tr>
<td>Jiazhen He</td>
<td>Measurement Theory in MOOCs via Topic Discovery</td>
</tr>
</tbody>
</table>

### Session 1B (Human Computer Interaction and Design)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hasan Shahid Ferdous</td>
<td>Technology at Mealtime: Beyond the 'Ordinary'</td>
</tr>
<tr>
<td>Melissa J. Rogerson and Martin Gibbs</td>
<td>Finding Time for Tabletop: Boardgame Play and Parenting</td>
</tr>
<tr>
<td>Fernando Estrada</td>
<td>Clinicians' perceptions of computerised support for mental health</td>
</tr>
<tr>
<td>Deepti Aggarwal</td>
<td>Understanding the Current Practices of Teleconsultation</td>
</tr>
</tbody>
</table>

### Session 2A (Machine Learning)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang Lei</td>
<td>Generalized Information Theoretic Cluster Validity Indices for Soft Clusterings</td>
</tr>
<tr>
<td>Mohadeseh Ganji, James Bailey and Peter Stuckey</td>
<td>Generalized Modularity for Community Detection</td>
</tr>
<tr>
<td>Zay Maung Maung Aye, Rao Kotagiri and Benjamin Rubinstein</td>
<td>Making Metric Learning Feasible for Large Datasets</td>
</tr>
<tr>
<td>Lida Rashidi, Sutharshan Rajasegarar and Christopher Leckie</td>
<td>Spectrum Preserving Graph Embedding</td>
</tr>
</tbody>
</table>
## Session Papers

### Session 2B (Business Information)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miguel Wood, Simon Milton and Graeme Shanks</td>
<td>Do extroverts Drive Commitment to Vote Behaviour?</td>
</tr>
<tr>
<td>Mohammad Moniruzzaman, Sherarh Kurnia and Sean Maynard</td>
<td>The Role of Business Intelligence in Supply Chain Agility</td>
</tr>
<tr>
<td>Diana Wong</td>
<td>Exploring the Use of Enterprise 2.0 and Its Impact on Social Capital within a Large Organisation</td>
</tr>
<tr>
<td>Carlos Andres Agudelo-Serna</td>
<td>Understanding Knowledge Leakage &amp; BYOD: From a Mobile Worker’s Perspective</td>
</tr>
</tbody>
</table>

### Session 3A (Computational Bioinformatics and Health Information Systems)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qingyu Chen, Karin Verspoor and Justin Zobel</td>
<td>Evaluation of A Machine-Learning based Duplicate-Detection Method for Bioinformatics Databases</td>
</tr>
<tr>
<td>Liyan Liu, Andrey Kan and Christopher Leckie</td>
<td>Comparative Evaluation of Performance Metrics for Shading Correction in Widefield Fluorescence Microscopy</td>
</tr>
<tr>
<td>Stephan Glöckner and Richard Sinnott</td>
<td>Web-based Feedback as a Tool for Data Quality Improvement in Biomedical Research</td>
</tr>
<tr>
<td>Seyed Mohammad Hossein Oloomi, Justin Zobel and Thomas C Conway</td>
<td>The Impact of Sequence Ambiguity in Read Mapping</td>
</tr>
</tbody>
</table>

### Session 3B (Distributed and Cloud Computing)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irum Bukhari, Aaron Harwood and Shanika Karunasekera</td>
<td>Optimum Benefit Protocol: A Fast Converging, Bandwidth-Efficient Decentralized Similarity Overlay</td>
</tr>
<tr>
<td>Safiollah Heidari and Rajkumar Buyya</td>
<td>Processing Large-Scale Graphs on Clouds</td>
</tr>
<tr>
<td>Xunyun Liu and Rajkumar Buyya</td>
<td>Auto Scaling within Apache Storm</td>
</tr>
<tr>
<td>Bowen Zhou</td>
<td>A Context Sensitive Offloading Scheme for Mobile Cloud Computing Service</td>
</tr>
</tbody>
</table>
### Session Papers

#### Session 4A (Natural Language Processing, Databases & Business Information)

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Duong</td>
<td>Neural Network Approach for Universal Dependency Parsing</td>
</tr>
<tr>
<td>Bahar Salehi</td>
<td>Let's Shoot the Breeze! Flexible Language Independent Multiword Expression Analysis</td>
</tr>
<tr>
<td>Maryam Fanaeepour and Benjamin I. P. Rubinstein</td>
<td>Differentially-Private Counting of GPS Points Regions in Facility Location Planning</td>
</tr>
<tr>
<td>Mohamad Rahimi Mohamad Rosman, Sherah Kurnia and Ivo Widjaja</td>
<td>Investigating the Determinants of Enterprise Content Management System (ECMS) Benefits</td>
</tr>
</tbody>
</table>

#### Session 4B (Optimization and Programming Languages)

<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maria Alejandra Rodriguez Sossa and Rajkumar Buyya</td>
<td>A Responsive Knapsack-based Algorithm for Resource Provisioning and Scheduling of Workflows in Clouds</td>
</tr>
<tr>
<td>Neelofar, Lee Naish and Rao Kotagiri</td>
<td>Multiple Bug Spectral Fault Localization Using Genetic Programming</td>
</tr>
<tr>
<td>David Clarke</td>
<td>Detecting Data Races in Java Programs</td>
</tr>
<tr>
<td>Andres Abeliuk</td>
<td>The Benefits of Social Influence in Cultural Markets</td>
</tr>
</tbody>
</table>
What do you think about that?

Tweet us @cis_dc and share your thoughts on the talks / presentations / posters you are listening to!

The hashtag to use is #cisdc2015

Find us on facebook: https://www.facebook.com/melbcisdc and give your valuable feedback about talks / presentations / posters you are listening to!
System U to Watson Personality Insights: The Transformation of A Research Project into a Commercial Offering

The measure of any successful researcher is the impact of their work. This impact can be measured scientifically through the number of citations of their published journal and conference papers, or, just as importantly in IBM Research, measured by the revenue it generates after being transformed into a commercial offering.

This presentation will introduce IBM Research - Australia and describe the transformation of a research project, called System U, into a commercial offering, called Watson Personality Insights.

Bio Sketch

Dr Anna Phan began her studies with a double degree in software engineering and science at The University of Melbourne, pursuing her competing interests in programming and science. Science won initially, and Anna continued studying particle physics and was involved in the ATLAS experiment at CERN that discovered the Higgs Boson. However, her interest in information technology triumphed and she is now a manager at IBM Research - Australia, leading the physical modelling and data analytics team.
## Databases

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differentially-Private Counting of GPS Points Regions in Facility Location Planning</td>
<td>18</td>
</tr>
<tr>
<td>Maryam Fanaeepour and Benjamin I. P. Rubinstein</td>
<td></td>
</tr>
</tbody>
</table>

## Business Information Systems

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do extroverts drive commitment to vote behaviour?</td>
<td>20</td>
</tr>
<tr>
<td>Miguel Wood, Simon Milton and Graeme Shanks</td>
<td></td>
</tr>
<tr>
<td>Investigating the Determinants of Enterprise Content Management System (ECMS) Benefits</td>
<td>21</td>
</tr>
<tr>
<td>Mohamad Rahimi Mohamad Rosman, Sherah Kurnia and Ivo Widjaja</td>
<td></td>
</tr>
<tr>
<td>The Role of Business Intelligence in Supply Chain Agility</td>
<td>22</td>
</tr>
<tr>
<td>Mohammad Moniruzzaman, Sherarh Kurnia and Sean Maynard</td>
<td></td>
</tr>
<tr>
<td>Exploring the Use of Enterprise 2.0 and Its Impact on Social Capital within a Large Organisation</td>
<td>23</td>
</tr>
<tr>
<td>Diana Wong</td>
<td></td>
</tr>
<tr>
<td>Understanding Knowledge Leakage &amp; BYOD: From a Mobile Worker’s Perspective</td>
<td>24</td>
</tr>
<tr>
<td>Carlos Andres Agudelo-Serna</td>
<td></td>
</tr>
</tbody>
</table>

## Computational Bioinformatics and Health Information Systems

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation of a machine-learning based duplicate-detection method for bioinformatics databases</td>
<td>26</td>
</tr>
<tr>
<td>Qingyu Chen, Karin Verspoor and Justin Zobel</td>
<td></td>
</tr>
<tr>
<td>Comparative evaluation of performance metrics for shading correction in widefield fluorescence microscopy</td>
<td>27</td>
</tr>
<tr>
<td>Liyan Liu, Andrey Kan and Christopher Leckie</td>
<td></td>
</tr>
<tr>
<td>Web-based Feedback as a Tool for Data Quality Improvement in Biomedical Research</td>
<td>28</td>
</tr>
<tr>
<td>Stephan Glöckner and Richard Sinnott</td>
<td></td>
</tr>
<tr>
<td>The Impact of Sequence Ambiguity in Read Mapping</td>
<td>29</td>
</tr>
<tr>
<td>Seyed Mohammad Hossein Oloomi, Justin Zobel and Thomas C Conway</td>
<td></td>
</tr>
</tbody>
</table>

## Distributed and Cloud Computing

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Responsive Knapsack-based Algorithm for Resource Provisioning and Scheduling of Workflows in Clouds</td>
<td>32</td>
</tr>
<tr>
<td>Maria Alejandra Rodriguez Sossa and Rajkumar Buyya</td>
<td></td>
</tr>
<tr>
<td>Optimum Benefit Protocol: A Fast Converging, Bandwidth-Efficient Decentralized Similarity Overlay</td>
<td>33</td>
</tr>
<tr>
<td>Irum Bukhari, Aaron Harwood and Shanika Karunasekera</td>
<td></td>
</tr>
</tbody>
</table>
### Paper Submissions

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing Large-Scale Graphs on Clouds</td>
<td>34</td>
</tr>
<tr>
<td>Safiollah Heidari and Rajkumar Buyya</td>
<td></td>
</tr>
<tr>
<td>Auto Scaling within Apache Storm</td>
<td>35</td>
</tr>
<tr>
<td>Xunyun Liu and Rajkumar Buyya</td>
<td></td>
</tr>
<tr>
<td>A Context Sensitive Offloading Scheme for Mobile Cloud Computing Service</td>
<td>36</td>
</tr>
<tr>
<td>Bowen Zhou</td>
<td></td>
</tr>
<tr>
<td><strong>Natural Language Processing</strong></td>
<td></td>
</tr>
<tr>
<td>Neural Network Approach for Universal Dependency Parsing</td>
<td>38</td>
</tr>
<tr>
<td>Long Duong</td>
<td></td>
</tr>
<tr>
<td>Bilingual Lexicon Induction from Small Quantities of Phonemic Transcriptions</td>
<td>39</td>
</tr>
<tr>
<td>Oliver Adams</td>
<td></td>
</tr>
<tr>
<td>Let’s Shoot the Breeze! Flexible Language Independent Multiword Expression Analysis</td>
<td>40</td>
</tr>
<tr>
<td>Bahar Salehi</td>
<td></td>
</tr>
<tr>
<td><strong>Human Computer Interaction and Design</strong></td>
<td></td>
</tr>
<tr>
<td>Technology at Mealtime: Beyond the ‘Ordinary’</td>
<td>42</td>
</tr>
<tr>
<td>Hasan Shahid Ferdous</td>
<td></td>
</tr>
<tr>
<td>Finding Time for Tabletop: Boardgame Play and Parenting</td>
<td>43</td>
</tr>
<tr>
<td>Melissa J. Rogerson and Martin Gibbs</td>
<td></td>
</tr>
<tr>
<td>Clinicians’ perceptions of computerised support for mental health</td>
<td>44</td>
</tr>
<tr>
<td>Fernando Estrada</td>
<td></td>
</tr>
<tr>
<td>Understanding the Current Practices of Teleconsultation</td>
<td>45</td>
</tr>
<tr>
<td>Deepti Aggarwal</td>
<td></td>
</tr>
<tr>
<td><strong>Machine Learning and Data Mining</strong></td>
<td></td>
</tr>
<tr>
<td>Tour Recommendation on Location-based Social Networks</td>
<td>48</td>
</tr>
<tr>
<td>Kwan Hui Lim</td>
<td></td>
</tr>
<tr>
<td>Privacy-Preserving Location Mining of Tweets</td>
<td>49</td>
</tr>
<tr>
<td>Shuo Wang and Richard Sinnott</td>
<td></td>
</tr>
<tr>
<td>Generalized Information Theoretic Cluster Validity Indices for Soft Clusterings</td>
<td>50</td>
</tr>
<tr>
<td>Yang Lei</td>
<td></td>
</tr>
</tbody>
</table>
# Paper Submissions

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized Modularity for Community Detection</td>
<td>51</td>
</tr>
<tr>
<td>Mohadeseh Ganji, James Bailey and Peter Stuckey</td>
<td></td>
</tr>
<tr>
<td>How Private Can I Be Among Public Users?</td>
<td>52</td>
</tr>
<tr>
<td>Elham Naghi Zadeh, James Bailey, Lars Kulik and Egemen Tanin</td>
<td></td>
</tr>
<tr>
<td>Modelling Pedestrian Activities in a Dynamic Environment</td>
<td>53</td>
</tr>
<tr>
<td>Tuan Doan, Sutharshan Rajasegarar and Christopher Leckie</td>
<td></td>
</tr>
<tr>
<td>Making Metric Learning Feasible for Large Datasets</td>
<td>54</td>
</tr>
<tr>
<td>Zay Maung Maung Aye, Kotagiri Ramamohanarao and Benjamin Rubinstein</td>
<td></td>
</tr>
<tr>
<td>Measurement Theory in MOOCs via Topic Discovery</td>
<td>55</td>
</tr>
<tr>
<td>Jiazhen He</td>
<td></td>
</tr>
<tr>
<td>Spectrum Preserving Graph Embedding</td>
<td>56</td>
</tr>
<tr>
<td>Lida Rashidi, Sutharshan Rajasegarar and Christopher Leckie</td>
<td></td>
</tr>
<tr>
<td><strong>Optimization and Programming Languages</strong></td>
<td></td>
</tr>
<tr>
<td>Multitle Bug Spectral Fault Localization Using Genetic Programming</td>
<td>58</td>
</tr>
<tr>
<td>Neelofar, Lee Naish and Rao Kotagiri</td>
<td></td>
</tr>
<tr>
<td>Detecting data races in Java programs</td>
<td>59</td>
</tr>
<tr>
<td>David Clarke</td>
<td></td>
</tr>
<tr>
<td>The Benefits of Social Influence in Cultural Markets</td>
<td>60</td>
</tr>
<tr>
<td>Andres Abeliuk</td>
<td></td>
</tr>
</tbody>
</table>
Poster Submissions

Tell Me What You Want And I Will Tell Others Where You Have Been
Anthony Quattrone, Elham Naghizade, Lars Kulik and Egemen Tanin
Towards a Painless Index for Spatial Objects
Rui Zhang, Jianzhong Qi, Martin Stradling and Jin Huan
Analysis of Road Traffic Using Contrast Mining and GPS Trajectories
Xiaoting Wang, Christopher Leckie and Tharshan Vaithianathan
Online Medical Consultation (OMC): A review of literature, practice, and development of an evaluation model
Ibrahim Al-Mahdi, Kathleen Gray and Reeva Lederman
Constraint Programming approach to the Steiner Tree Problem with side constraints using learning
Diego de Uña, Graeme Gange, Peter Schachte and Peter J. Stuckey
Exploring Factors that Impact on Saudi Female International Student’s use of Social Technologies as an Information Source
Haifa Binsahl, Shanton Chang and Rachelle Bosua
Tour Recommendation on Location-based Social Networks
Kwan Hui Lim
The Impact of Big Data Analytics on Information Security Process Performance
Humza Naseer, Sean Maynard, Graeme Shanks and Atif Ahmad
Effective trust and reputation strategies
Sergio Freschi, Michael Kirley and Tim Miller
Communication in Artificial Agent Teams for Interdependent Tasks
Ronal Singh
Better Health Explorer
Patrick Cheong-Iao Pang, Shanton Chang, Karin Verspoor and Jon Pearce
Text Categorisation using Convolutional Neural Networks
Fei Liu, Timothy Baldwin and Trevor Cohn
Towards a Model of Information Security Management Practices in Organisations
Moneer Alshaikh
# Poster Submissions

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Base Population over BioMedical Data</td>
<td>75</td>
</tr>
<tr>
<td>Nagesh Panyam Chandrasekaraaswy, Rao Kotagiri and Trevor Cohn</td>
<td></td>
</tr>
<tr>
<td>Natural Language Coreference Resolution for Biomedical Event Extraction</td>
<td>76</td>
</tr>
<tr>
<td>Joo-Young Choi, Karin Verspoor and Justin Zobel</td>
<td></td>
</tr>
<tr>
<td>The voluntary use of devices implanted in and through the skin</td>
<td>77</td>
</tr>
<tr>
<td>Kayla Heffernan, Frank Vetere and Shanton Chang</td>
<td></td>
</tr>
<tr>
<td>Using Community Question-Answering Data to Answer User’s Questions</td>
<td>78</td>
</tr>
<tr>
<td>Doris Hoogeveen, Tim Baldwin and Karin Verspoor</td>
<td></td>
</tr>
<tr>
<td>Query Processing on Time-evolving Graphs</td>
<td>79</td>
</tr>
<tr>
<td>Masoomeh Zameni, Christopher Andrew Leckie and Masud Moshtaghi</td>
<td></td>
</tr>
<tr>
<td>Implementing the greedy Set Cover algorithm efficiently for disk-resident data</td>
<td>80</td>
</tr>
<tr>
<td>Ching Lih Lim, Alistair Moffat and Anthony Wirth</td>
<td></td>
</tr>
<tr>
<td>Is greed good? Towards sustainable and profitable resource harvesting</td>
<td>81</td>
</tr>
<tr>
<td>Friedrich Burkhard von der Osten</td>
<td></td>
</tr>
<tr>
<td>Nyumbani: Designing Technologies for Indigenous Knowledge</td>
<td>82</td>
</tr>
<tr>
<td>Kagonya Awori</td>
<td></td>
</tr>
<tr>
<td>Understanding Knowledge Leakage &amp; BYOD: From a Mobile Worker’s Perspective</td>
<td>83</td>
</tr>
<tr>
<td>Carlos Andres Agudelo Serna</td>
<td></td>
</tr>
<tr>
<td>The Fuzziness of Next Generation Sequencing Analysis: A Case Study on the Need for Reproducibility of Results</td>
<td>84</td>
</tr>
<tr>
<td>Sehrish Kanwal, Andrew Lonie, Richard Sinnott and Charlotte Anderson</td>
<td></td>
</tr>
<tr>
<td>Speeding Up Visual Field Tests By Incorporating Spatial Models</td>
<td>85</td>
</tr>
<tr>
<td>Nikki Rubinstein, Allison McKendrick and Andrew Turpin</td>
<td></td>
</tr>
<tr>
<td>Appropriation of Social Media for Knowledge Sharing By Patients With Chronic Illnesses</td>
<td>86</td>
</tr>
<tr>
<td>Nwakego Isika, Antonette Mendoza and Rachelle Bosua</td>
<td></td>
</tr>
<tr>
<td>Technology at Mealtime: Beyond the ‘Ordinary’</td>
<td>87</td>
</tr>
<tr>
<td>Hasan Shahid Ferdous</td>
<td></td>
</tr>
</tbody>
</table>
Poster Submissions

DesTeller: A System for Destination Prediction Based on Trajectories with Privacy Protection  88
Andy Yuan Xue, Rui Zhang, Yu Zheng, Xing Xie, Jianghui Yu and Yong Tang

Improving Data Quality in Biomedical Trials through Automating Feedback  89
Stephan Gloeckner, Anthony Stell, Jemie Effendy, Irina Bancos, Wiebke Arlt and Richard Sinnott

Robust Inferences of Travel Paths from GPS Trajectories  90
Hengfeng Li, Lars Kulik and Rao Kotagiri

Value-driven approach for Requirements Engineering in People-Oriented Software  91
Mohammadhossein Sherkat, Tim Miller and Antonette Mendoza

Structural Alignment as the Basis for Significant Changes Detection between Versioned Sentences  92
Ping Tan, Karin Vespoor and Timothy Miller

Distributed stream clustering using micro-clusters on Apache Storm  93
Pasan Karunaratne, Shanika Karunasekera and Aaron Harwood

Summarisation Strategies for Symbolic Execution  94
Yude Lin

Exploiting Text and Network Context for Geolocation of Social Media Users  95
Afshin Rahimi, Duy Vu, Trevor Cohn and Timothy Baldwin

Tensions in Digitising Boardgames  96
Melissa J. Rogerson

Word-level Bit Propagation on fixed-width integers  97
Wenxi Wang

Optimisation Modelling for Software Developers  98
Kathryn Francis, Peter J. Stuckey and Sebastian Brand

Using Map Information to facilitate Activity Recognition on Smart-phone  99
Weihao Cheng, Rao Kotagiri and Rui Zhang

Enhancing Waiting Experiences in Video Mediated Communications  100
Behnaz Rostami Yeganeh
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designing Improvised Coordination in Agent Organizations</td>
<td>101</td>
</tr>
<tr>
<td>Kathleen Keogh</td>
<td></td>
</tr>
<tr>
<td>Designing Technology for Human-Animal Encounters at the Zoo</td>
<td>102</td>
</tr>
<tr>
<td>Sarah Webber</td>
<td></td>
</tr>
<tr>
<td>Voice and the virtual experience</td>
<td>103</td>
</tr>
<tr>
<td>Fraser Allison</td>
<td></td>
</tr>
<tr>
<td>Sequencing Operator Counts</td>
<td>104</td>
</tr>
<tr>
<td>Toby Davies</td>
<td></td>
</tr>
<tr>
<td>Transforming Situation Calculus Histories for Efficient Reasoning</td>
<td>105</td>
</tr>
<tr>
<td>Christopher Ewin</td>
<td></td>
</tr>
<tr>
<td>Founded Bounds and Model Counting for Answer Set Programming</td>
<td>106</td>
</tr>
<tr>
<td>Rehan Abdul Aziz</td>
<td></td>
</tr>
<tr>
<td>Information Security Strategy in Organisations</td>
<td>107</td>
</tr>
<tr>
<td>Craig Horne</td>
<td></td>
</tr>
</tbody>
</table>
Databases
ABSTRACT
Due to the high demand for location-based services (LBSs), large spatio-temporal datasets of moving objects are being created every day. An important task in spatial data analytics is to service range queries by returning moving object counts within a queried region. In some applications, third parties are interested in the number of users’ GPS regions, hot spots, to set up facility services. As a result, the question of how to keep a user’s data private whilst enabling spatial data analytics by third parties has become an urgent research direction. In this extended abstract, we utilize the benefit of aggregates, in which work with individuals’ data are not required, and apply the strong notion of differential privacy. In other words, we discuss differentially private counting of planar bodies, with the application of finding hot spots based on users’ GPS regions.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Spatial databases and GIS; D.4.6 [Security and Protection]: Privacy

Keywords
Aggregate Data; Count Information; Differential Privacy; Distinct Counting Problem; Euler Histograms; Location Privacy; Spatial Databases; Spatial Data Analytics

1. INTRODUCTION
The proliferation of location-based services (LBSs) has led to an ever-increasing amount of spatio-temporal data, and a growing need to process and analyse this data. A wide range of applications benefit from spatio-temporal data such as: finding a service in the vicinity of a user’s location; determining the best site to set up a new facility such as a café, i.e., facility location planning; and traffic planning by monitoring the number of vehicles in a specific region [2]. However, significant concerns regarding an individual’s data privacy can limit adoption [3]. Even with only GPS trajectories, it is possible to track and identify a person via a reverse lookup inference attack [5]. In particular, many trips start or end at a user’s persistent location e.g., home or work. Therefore, a trajectory ID is nearly as revealing as a person’s ID. We propose differentially private counting of planar bodies using a post-processing approach and release Euler histogram counts [4] in a differentially-private manner, which we discuss next.

2. OUR APPROACH
For the case of releasing Euler histogram counts of moving objects, the service provider may be required to guarantee a strong notion of privacy: that the released counts reveal nothing significant about users not already known to a powerful attacker. Such a requirement is the focus of differential privacy. One solution could be adding Laplace noise to each vector component consisting of all face, edge, and vertex counts. However, the amount of noise which is added, will be quite large, depending on the region size, grid cell size, and consequently the number of grids. In that, if we change one record in the dataset all the histogram counts could change in the worst case, therefore the amount of noise that we need to add will be $O(n \times n)$, where $n$ is the number of grid cells. This leads to a high level of inaccuracy. Therefore, the main challenges to be faced here are accuracy as well as consistency as we cannot merely add noise independently. In our approach to tackle this problem, we assume that we have convex hull of GPS points (e.g. GPS points regions) with bounded diameter, thus the amount of noise to be added, will be smaller and under control. In addition, we take a post-processing approach: we have the original dataset of histograms, and we also have the noisy one via applying the computed noise as above. Thereby, we take a linear programming approach, calculating a new set of Euler histogram counts which minimizes the distance between the noisy set and the third set, subject to our constraints. Finally, we round the third set of counts. By constraint on our responses, we achieve consistency and consequently accuracy [1] as well as hiding the fact that we are adding noise to the responses.

3. REFERENCES
Business Information Systems
Do extroverts drive commitment to vote behaviour?

[Extended Abstract]

Miguel Wood
University of Melbourne
mwood3960@gmail.com

Simon Milton
University of Melbourne
simon.milton@unimelb.edu.au

Graeme Shanks
University of Melbourne
gshanks@unimelb.edu.au

ABSTRACT
Across Western societies political participation is in decline. Political advocacy however is undergoing rapid transformation led by a disruptive wave of IT led innovation. Campaigns are increasingly fought online with recent advances leveraging online social networks and behavioural science to target and influence voters. But is the likelihood and distribution of political participation uniform across social networks. In this paper, we examine the interaction of personality traits, online social networks and human social motives to enact social influence within a community during an election. Our research model specifically investigates whether extroverts drive commitment to vote behavior using an online vote plan. The insight may help political campaigns to design and optimize future behavioural targeting strategies that activate social network effects to achieve campaign goals along with offering a new pathway for rejuvenating democracy.

Categories and Subject Descriptors
H.1.1 [Systems & Information Theory]: Information theory; H.1.2 [User/Machine Systems]: Human factors, software psychology, social networks; J.4 [Social & Behavioural Science]: Sociology

Keywords
Personality traits, political efficacy, online social networks, implementation intentions, political campaigns

1. INTRODUCTION
Investment and increased maturity of development operations, predictive analytics and online social networks has yielded greater precision in voter targeting and field operation effectiveness for political campaigns. Next generation capabilities are oriented towards enhanced social targeting, real-time analytics to overtake polling, and genuine digital integration.

The migration of political campaigning into the digital realm has revealed important effects where social networks and social pressure are combined to increase voting and also differences in the online political behaviour of extroverted users [1, 2].

Personality traits and political self-efficacy operate in concert to account for political behaviour [3]. Personality traits are enduring dispositions that explain behaviour and preferences. Political self-efficacy is the belief that individual political action does have, or can have, an impact upon the political process. The existing literature remains dominated by the individual focused paradigm and silent on the contribution of traits and self-efficacy of political active individuals in influencing their social relations.

The integration of personality, social networks and human social motives [4] presents an opportunity to understand whether extroverts shape the political attitudes and behaviours of those around them.

2. RESEARCH MODEL
The research model is based on the theoretical framework of personality traits, internal political efficacy (social cognitive theory), and social influence and contagion. The research model is represented in two parts: A and B; comprising six constructs and six hypotheses. Part A is concerned with the role of personality traits in shaping political beliefs and indirectly political participation in the form of an online vote plan derived from the psychological construct of implementation intention (“The Belief Model”)

![Figure 1: The Belief Model](image1.png)

Part B traces extroverted individuals’ with high political efficacy that make an early commitment to vote and the influence they exert (via push notification) on their online social networks and the human social motivations which drive adoption of similar behaviour in their social circle of influence (“The Behaviour Model”)

![Figure 2: The Behaviour Model](image2.png)

3. REFERENCES

Business Information Systems
ABSTRACT
Organisations implement Enterprise Content Management System (ECMS) to yield specific benefits. This research attempts to investigate determinants that drive achievement of ECMS benefits. Research context and knowledge gaps are drawn up and the aim of the study is highlighted along with research questions and methodology. The paper is then concluded with expected contributions from this study.

Categories and Subject Descriptors
K.4.3 [Organizational Impacts]: Automation;

Keywords
Enterprise content management, enterprise content management system, content management, information system benefits.

1. INTRODUCTION
Organisations implement information system (IS) to yield benefits from its usage. ECMS was initiated to help organisations manage its information assets in digital form (text, graphic, email, 3D data, etc.). A glance into ECM literature shows that little research has been conducted in terms of investigating the benefits of ECMS for organisations (Nordheim and Paivarinta, 2006, Grahlmann et al., 2010). Therefore, this research will attempt to investigate the benefits that organisations seek from the use of ECMS and factors that lead them into the receipt of benefits.

2. RESEARCH CONTEXT
The use of content in digital form is incrementing at an incredible speed and goes beyond organisation control. Over 2.5 quintillion bytes of data have been created in the past two years alone. Digital website content has shown a tremendous growth rate, increasing from nearly 1,000 pages in the early 1990s into over two billion webpages in 2001 and has reached a new milestone of one billion from nearly 1,000 pages in the early 1990s into over two billion webpages in 2001 and has reached a new milestone of one billion. The rapid growth of digital content has led to a phenomenon called “content chaos”. It has caused organisations to be deprived of their information assets. Hence, ECMS was initiated to help manage digital content.

3. KNOWLEDGE GAPS
Researchers argue that: (1) organisations are struggling to realise benefits from ECMS; (2) ECMS is underutilised due to the fact that the system does not align with the way businesses work and lack of user training; and (3) the post-implementation stage in ECMS is completely neglected and requires much attention from academics and practitioners.

4. AIM AND RESEARCH QUESTIONS
Therefore, the aim of this study is twofold: (1) to investigate the benefits that drive organisations towards ECMS implementation; and (2) exploring the determinants that drive the achievements of ECMS benefits. Our proposed research questions are as follows:
RQ1: What are the benefits of ECMS use?
RQ2: What are the factors that drive the achievements of ECMS benefits?

5. PROPOSED METHOD
In an attempt to answer RQ1, a multiple case study will be adopted based on Yin (2013). Two organisations will be selected and an interview will be conducted with two members from each organisation, supported by observation and reviewing existing documentation. For RQ2, a research model will be developed, tested and validated based on quantitative study using surveys. The population will be selected from Malaysian organisations that have used ECMS for more than two years. The instrument will undergo a pilot test in order to ensure its validity and Cronbach’s alpha will be used to test the reliability of the questionnaires. The data will be recorded using Statistical Package for Social Science (SPSS) version 23 and analysed.

6. EXPECTED STUDY CONTRIBUTIONS
The expected study contributions are threefold: Firstly, it hopes provide a foundation for academics and practitioners for making strategic decisions to adopt ECMS. Secondly, it aims to enrich understanding on how organisations can yield benefits during the post-implementation stage. Thirdly, we introduce ECMS benefits framework to help organisations justify their investment into ECMS.

7. ACKNOWLEDGMENTS
The author would like to thanks Universiti Teknologi MARA (UiTM) and Kementerian Pendidikan Malaysia (KPM) for the Skim Latihan Akademik IPTA (SLAI) sponsorship.

8. REFERENCES
The Role of Business Intelligence in Supply Chain Agility

Mohammad Moniruzzaman
Department of Computing and
Information Systems
monirm@student.unimelb.edu.au

Dr. Sherah Kurnia
Department of Computing and
Information Systems
sherahk@unimelb.edu.au

Dr. Sean Maynard
Department of Computing and
Information Systems
sean.maynard@unimelb.edu.au

ABSTRACT

This study explores how the use of Business Intelligence (BI) in Supply Chain management (SCM) can improve supply chain (SC) agility. We seek to expand the current understanding of the use of BI to support evidence-based decision making in the main process areas of SCM (Plan, Source, Make, Deliver and Return). We also want to investigate how BI contributes to SC agility of an organization. The result of the study will help organizations implement BI effectively to support SCM and improve SC agility.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; H.4.2 [Decision Support].

Keywords
Supply Chain Management, Agile, Business Intelligence.

1. INTRODUCTION

SC plays a vital role in an organisation to survive in a dynamic market condition. SC needs to be agile enough to convert market changes into opportunities. However, how agility in SC can be achieved by organisation is unclear.

2. SCM AND SC AGILITY

SC is a network of organisations and activities for raw material acquisition, production and distribution of final products/services to satisfy end user needs in the market. Supply chain management (SCM) is the process of planning, implementing and controlling the efficient, cost-effective flow and storage of raw materials, in-process inventory, finished goods, and related information for the purpose of conforming to customer requirements (Lambert et al., 1996). For organisations to stay competitive, the SC needs to be agile, which requires effective SCM. Achieving an agile SC is challenging for organisations because of the complexity involved in managing SC activities cross organisational boundaries (Soni et al., 2013).

3. RESEARCH AIM AND QUESTION

Many organisations have deployed Information Technologies to achieve agile SC. One such technology, Business Intelligence (BI), enables organisations to analyse integrated organisational data to support various evidence-based decision making processes. Since SCM is complex, a better understanding of how BI helps organisations achieve SC agility requires in-depth studies of each major area of SCM. This study explores how SC agility is achieved through the use of BI in SCM. The research question for this study is: ‘How does BI-enabled SCM improve the agility in Supply Chain?’

4. PROPOSED RESEARCH MODEL

Propositions P1a-P5a in proposed research model (figure: 1) suggest that the greater the BI capability, the greater the evidence-based decision making in respective SCM areas while propositions P1b-P5b suggest that the greater the evidence-based decision making in respective SCM areas, the greater the agility in supply chain.

Figure 1: Proposed Research Model

4.1 Key Constructs

BI Capability: The ability of an organization to integrate multiple data sources and present summarized information in the form of graphs, charts, trends and other user-friendly formats to provide insight to users for decision making.

Evidence-based SCM: Management of plan, source, make, deliver and return processes of the SC based on facts derived from internal/external data in an organization.

Agility in SC: The capability of a SC to deal with dynamic market conditions. Agility in SC can be characterized as Flexibility, Quickness, Responsiveness and Competency.

5. RESEARCH METHOD

A qualitative, multiple case study research method involving multiple participants from 5 organisations will be used for this study. Within each organisation, we will interview 6-8 SCM personnel. Each participant will be asked a set of closed-ended and open-ended questions regarding how BI supports decision making within each area of SCM.

6. REFERENCES


Exploring the Use of Enterprise 2.0 and Its Impact on Social Capital within a Large Organisation

[Extended Abstract]
Diana Wong
University of Melbourne
d.wong1@student.unimelb.edu.au

ABSTRACT
Despite the rampant adoption of Enterprise 2.0, there is lack of empirical evidence of how Enterprise 2.0 is aptly supporting the business objectives. Social capital theory will be used as a theoretical lens to understand the impact and implications of individual use of Enterprise 2.0. A qualitative case research was conducted and the result delineates the areas of impact on the social capital through the individual use of Enterprise 2.0.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous;

Keywords
Enterprise Social Networking, Enterprise 2.0, Social Capital,

1. INTRODUCTION
Enterprise social networking platform, or Enterprise 2.0, is a platform which consists of a wide range of integrated social media tools (Li et al., 2012). The use of Enterprise 2.0 enables individuals to connect to each other and utilize the integrated social media tools in an activity stream supporting business’ objectives. Many organisation have adopted Enterprise 2.0, however, however, the potential impact of Enterprise 2.0 is still not well understood. Therefore, the research postulates the use of social capital theory which explains that the network connections between individuals within an organisational facilitate mutually benefit collective actions. Thus, to address the identified gaps, the central research question is as follow: *How does the individual use of Enterprise 2.0 impact the social capital within a large organisation?*

Up until now, the potential of value created from the use of Enterprise 2.0 on its social capital within the organization is not fully explored. The synthesis of literatures[2], [3] indicates the lack of systematic and holistic explanation of how Enterprise 2.0 impacts the social capital. Drawing on Nahapiet and Ghosal’s [4] research, the authors view social capital as “the sum of actual and potential resources embedded within, available through, and derived from the network of relationships possessed by an individual or social unit.” The authors categorized the understanding of social capital into structural, cognitive, and relational dimension.

2. RESEARCH METHOD
To ascertain the impact from the use of Enterprise 2.0 on the various dimensions of social capital, a single in-depth qualitative case study was conducted with a large professional services organisation that has multiple offices worldwide and within Australia. The organisation adopted Yammer as its Enterprise 2.0 platform. 12 interview participants were recruited to gather their behaviours, perspectives, feelings, and experiences using Enterprise 2.0.

3. FINDINGS AND DISCUSSION
The study examines both the individual and collective social capital and outlines the following:
1. Structural Dimension: Based on the findings, the network of the organisation is formed through vertical and lateral association of individuals and groups. The varying strength of the social network ties creates opportunities, provide accessibility, and novelty of knowledge across the network of individuals, positively impacting the social capital within the organisation.
2. Cognitive Dimension: The use of Yammer increases individual’s awareness and understanding of the work group, organisation, and business, leading to serendipitous opportunity. The impact on the individual social capital comprises the mobilization of the available resources within the social network – structural dimension, combined with the actions and outcomes which are intertwined with aspects of collective social capital.
3. Relational Dimension: The use of Yammer increases individual’s awareness and understanding of the work group, organisation, and business, leading to serendipitous opportunity. The impact on the individual social capital comprises the mobilization of the available resources within the social network – structural dimension, combined with the actions and outcomes which are intertwined with aspects of collective social capital.

4. CONCLUSION
Enterprise 2.0 provides the avenue for social networking and interaction which fosters strengthening of relationships leading to development of social capital within large organisations.

5. REFERENCES
Understanding Knowledge Leakage & BYOD: From a Mobile Worker’s Perspective

[Extended Abstract]

Carlos A. Agudelo S
University of Melbourne
cagudelo@student.unimelb.edu.au

Rachelle Bosua
University of Melbourne
rachelle.bosua@unimelb.edu.au

Atif Ahmad
University of Melbourne
atif@unimelb.edu.au

Sean Maynard
University of Melbourne
Seanbm@unimelb.edu.au

ABSTRACT
This research examines the phenomenon of knowledge leakage through the adoption of BYOD (Bring Your Own Device) by organizations and among mobile knowledge workers and its impacts on organizational information security.

Categories and Subject Descriptors
K.6 [Management of Computing and Information Systems]: K.6.5 [Security and Protection]: Miscellaneous; K.6.m Security

Keywords
BYOD, Bring your own device, Knowledge leakage, Knowledge protection, Mobility, Information security.

1. INTRODUCTION
Organizational knowledge is a corporate asset that provides sustainable competitive advantage to organizations in today’s marketplace where innovation and knowledge based resources play a paramount role in the strategy and future of their performance. However, paradoxically while knowledge is considered an asset that must be protected, knowledge managers still seem to pay little attention to security issues. Furthermore, with the use of pervasive mobile computing, knowledge workers have become more mobile, increasing the risks of information security incidents and knowledge leakage when performing their knowledge work outside organization’s secure boundaries, i.e., mobile contexts. The problem exacerbates even more with the adoption of BYOD (Bring your own device) policies by organizations, as it allows knowledge workers use their own devices in the hope of increasing productivity and in the detriment of information security. By acknowledging the dynamic nature of knowledge leakage and BYOD, it is important to understand the factors to drive knowledge workers to engage in contexts that drastically enhances the risks of knowledge leakage.

2. MOTIVATION
The extant literature about knowledge management focuses on knowledge sharing neglecting knowledge protection as it is considered a barrier to prevent knowledge from being creating and flowing in organizations. Despite the incidents of corporate espionage and knowledge leakage in recent years, the research on knowledge protection is in its infancy and further investigation is necessary to address this gap. Moreover, the focus on the scant literature addresses the dimensions of organizational product and processes which corresponds to explicit knowledge while neglecting people, which is the most important resource in a knowledge intensive environment.

In addition to this, BYOD brings other security challenges to the picture, as it allows mobile knowledge workers to be outside organizational secure boundaries and expose the organizational knowledge to the outside world.

Furthermore, in the current literature these issues are approached from a rather technical perspective neglecting the human dimension to the problem.

Thus, the study of behavioral attitudes which drive mobile knowledge workers must be investigated not only because they are one of the avenues whereby knowledge leakage transpires but also because the tacit knowledge embedded in knowledge workers is not being addressed and it is costing organizations its competitive advantage, i.e., valuable knowledge assets and resources.

3. RESEARCH AIM AND QUESTIONS
This research aims to provide empirical data through a systematic investigation on the BYOD (Bring your own device) adoption among mobile workers which leads to organizational knowledge leakage and the strategies used by organizations to prevent this problem. In order to achieve the aims of this research below are the research questions that this study seeks to answer:

- What factors contribute to mobile knowledge workers’ leaking organizational knowledge unintentionally through the use of BYOD in various contexts?
- How can organizations protect their organizational knowledge from being leaked by mobile workers using BYOD?

4. REFERENCES
Computational Bioinformatics
and Health Information Systems
Evaluation of a Machine Learning Duplicate Detection method for Bioinformatics Databases

[Extended Abstract]

Qingyu Chen  Justin Zobel  Karin Verspoor  
Department of Computing and Information Systems  
University of Melbourne, Parkville, Australia, 3010  
qingyuc1@student.unimelb.edu.au, jzobel,karin.verspoor@unimelb.edu.au

ABSTRACT

We reproduce and evaluate a previously published duplicate detection method for bioinformatics databases that is based on inference of association rules. Our experiments indicate that the method is not promising for broad applicability in genomic sequence databases.

1. INTRODUCTION

A duplicate in a bioinformatics database is an instance where two (or more) records represent the same biological entity. Such records are typically not identical, but may be redundant (with minor differences), incomplete, or inconsistent. While the presence of duplicates has been regarded as a key data quality issue since the early 1990s, neither the extent nor the type of duplicates has been thoroughly investigated. Nor has there been extensive investigation of discovery methods, in contrast with the start of art in general domains like text retrieval. In particular, machine learning techniques are used in other contexts for discovery of duplicates.

We have replicated a published method based on machine learning [2] and evaluated it against a new collection of submitter-labelled duplicates in nucleotide databases. In our results, the best rule in the original study can only detect 0.2% of the duplicates. The rules mined from the collection also have poor performance; many have less than 1% precision and high false negative rates.

2. METHOD AND RESULT

The replicated method consists of four steps. First, record fields are selected for comparison. Second, similarity of selected records is computed using a variety of techniques, such as edit distance and Boolean matching. Third, association rules are mined from the similarity scores. Fourth, generated rules are applied for detection of duplicates in a dataset. We replicated this method by using the model and techniques from the original work; in some cases, where material was unclear, we made reasonable assumptions.

Our test dataset contains 3,998 groups of duplicate or merged records in primary nucleotide databases such as GenBank [1]. We first examined how many duplicates in the test dataset can be found using the best rule as determined in the original study. We then quantified the performance of new rules generated from the test dataset. Results of the new rules are shown in Figure 1.

Figure 1: Non-promising results for rules with support of more than 50% (left half) or less than 50% (right half, shown labelled in grey).

Neither of the evaluations provides promising results. The best rule in the original study can only find 0.2% (17 out of 7,105) duplicates in the test collection. Figure 1 also illustrates that none of rules derived from the test dataset has reasonable performance, and there is no obvious difference between the rules with more than 50% support (the proportion of an itemset over the training dataset) and those with less. We are confident that the method is correctly implemented, and thus find that its performance is extremely poor.

3. REFERENCES

Comparative evaluation of performance metrics for shading correction in widefield fluorescence microscopy

[Extended Abstract]

Liyan Liu
Department of Computing and Information System
liyanl@student.unimelb.edu.au

Andrey Kan
Walter and Eliza Hall Institute of Medical Research
akan@csse.unimelb.edu.au

Christopher Leckie
Department of Computing and Information System
caleckie@unimelb.edu.au

ABSTRACT
Widefield fluorescence microscopy is a valuable technology, but it suffers from the inherent problem of intensity inhomogeneity due to uneven illumination, also known as the shading artefact. This will lead to inaccurate estimates of single-cell features such as average and total intensity. Numerous shading correction methods have been proposed to remove this effect. In order to compare the performance of different methods, many quantitative performance metrics have been developed. However, there is little discussion about which performance metric should be generally applied for evaluation on real data, where the ground truth is absent. Thus, we aim to address the problem for shading correction performance evaluation in the absence of ground truth.

Categories and Subject Descriptors
I.4 [Image processing and computer vision]: General; J.3 [Computer Application]: Life and medical sciences—Medical information systems

Keywords
Widefield fluorescence microscopy, shading correction methods, quantitative performance metrics

1. INTRODUCTION
Widefield fluorescence microscopy coupled with single cell tracking is an increasingly important tool in cell biology [2]. However, the acquired images are usually corrupted by shading, which may cause an inaccurate measurement of the average intensity of the cells. As is shown in Figure 1, we can clearly see that it is much darker in the boundaries than elsewhere. Besides, many image processing applications such as segmentation, tracking or quantitative analysis are highly sensitive to the shading artefact [2]. Hence, it is essential to eliminate the shading effect beforehand.

In order to show whether one correction method is feasible for fluorescence images and superior to other methods, numerous quantitative metrics have been proposed. However, there is no consensus regarding which performance metric should be applied in the absence of ground truth. Furthermore, little work has been done to evaluate and compare the performance of different shading correction methods, which makes it difficult to choose the best correction method. Thus, it is significant to compare all the updated shading correction approaches and determine an appropriate one. Therefore, our purpose is to determine a performance metric and select an optimal correction method.

2. METHOD
Since it is proved that the evaluation results vary with different quantitative performance metrics, we need to determine a reasonable performance metric for the evaluation on fluorescence images [1]. When the ground truth is available, it is generally accepted that the normalized $L_2$-norm of the difference between real uniform image and the estimated uniform image is a reliable quality measure [1]. Hence, we first perform experiments on synthetic images to determine a reasonable metric, which is consistent with $L_2$-norm and does not rely on ground truth. Next we apply this metric in real data and compare the performance of different methods.

3. CONCLUSIONS
In summary, our contributions are as follows: 1) address the problem of evaluation in the absence of ground truth; 2) select a best correction method for fluorescence images from the previous established methods. If the optimal method could not be achieved, we aim to propose a novel approach.

4. REFERENCES
Web-based Feedback as a Tool for Data Quality Improvement in Biomedical Research

[Extended Abstract]

Stephan Glöckner
Department of Computing and Information Systems, University of Melbourne
sglockner@student.unimelb.edu.au

ABSTRACT

International biomedical registries have to be driven by quality to ensure rapid translation of findings in basic science to the clinical domain. Unfortunately, retrospective data cleaning is often performed at the end of trials and studies rather than through ongoing quality assurance. This research hypothesis that automated data quality feedback can tackle this problem and improve the quality of data for improved studies and ultimately translation to clinical management.

Categories and Subject Descriptors
K.6.4 [System Management]: Quality assurance; J.3 [Life and Medical Science]: Medical information systems

Keywords
Data Quality Feedback, Biomedical Informatics, Data Entry

1. INTRODUCTION

Clinical data management systems developed for biomedical studies rarely perform quality assurance procedures during ongoing trials. Unfortunately more often than not, data cleaning after a trial is the traditional approach however this is often too late. Furthermore to handle the increasing amount of information captured in a clinical trials, heterogeneous data management systems are needed since weaknesses in data processes can lead to data inconsistencies and make the vision of rapid knowledge transfer and personalized medicine in the post-genomic age, a challenge or indeed impossible. International biomedical registries have to be driven by quality and validation processes need to be applied on datasets that are collected [1]. For most trials, data is manually collected and entered into targeted web forms - electronic case report forms (eCRFs).

2. PROPOSED CONCEPT

We suggest automated data quality feedback in combination with improvement suggestions throughout the course of a trial can be used to improve data capture and the usefulness of the collected datasets for the scientist. We classify feedback into real-time feedback during the data entry, investigator feedback for monitoring and training purposes, and community feedback (see Figure 1).

3. CASE STUDY

To measure the effectiveness of this concept, we applied investigator and community feedback on an ongoing multicenter clinical trial. We calculated data completeness and record eligibility as part of the data accuracy for this trial. The eligibility score included all from principal investigator defined compulsory items for the trial objectives. We discovered that around 40% of all trial record didn’t meet these requirements. We reported this eligibility and overall data quality to every center over 5 months and reached an overall quality improvement of 13% [2].

4. CONCLUSIONS

To develop a general framework for automated feedback, the suggested concept must be shown to work across different communities. To this end we show the way in which a range of biomedical projects have been designed with data quality and feedback mechanisms to increase the usefulness of the captured information.

5. ACKNOWLEDGMENTS

Our thanks to ESF-ENSAT for funding this project and to Mr. Anthony Stell and Jemie Effendy for the software development.

6. REFERENCES


The Impact of Sequence Ambiguity in Read Mapping

[S Mohammad H Oloomi]
Department of Computing and Information Systems,
The University of Melbourne
soloomi@student.unimelb.edu.au

[Thomas C Conway]
IBM Research, Australia
tconway@au1.ibm.com

[Justin Zobel]
Department of Computing and Information Systems,
The University of Melbourne
jzobel@unimelb.edu.au

ABSTRACT
Ambiguity, when a read can be mapped to more than one location, is one of the important challenges of the process of mapping reads to reference sequence. Despite the importance of ambiguity resolution in read mapping it has been less investigated separately to date. In this article, we first review the sources of ambiguity together with the approaches taken by existing mapping methods to handle ambiguous reads. We then investigate the impact of ambiguity in read mapping.

Categories and Subject Descriptors
J.3 [Life and Medical Sciences]: Biology and genetics—Computational Biology and Bioinformatics

Keywords
Read mapping, ambiguity, mapping accuracy, next generation sequencing, DNA sequencing

1. INTRODUCTION
Finding the genomic and proteomic sequences of the samples being studied is an essential step in many biological research areas. In Next Generation Sequencing methods, DNA is sheared randomly into numerous small fragments. These fragments are then sequenced in parallel, producing small sequences of DNA, called reads.

A previously sequenced DNA can often be used as a reference sequence for constructing the sequence of future samples from the same (or closely related) species. The process of constructing the sequence by mapping reads to a reference genome, is described as read mapping (or alignment).

One of the important challenges for achieving accurate mappings is sequence ambiguity. Ambiguity occurs when a read can be mapped to more than one location on the reference sequence. There are five factors causing ambiguity in read mapping: repetitive regions in the genome, genetic variability, sequencing errors, mapping limitations, and multiple-organism samples.

There are four approaches used by existing read mapping tools for handling ambiguity in read mapping: discarding ambiguous reads, reporting the best mapping, reporting all mappings and manually resolving ambiguous reads.

Making the right decision when facing ambiguity in mapping reads could be vital for the downstream analyses. For example, it has been shown that ignoring all repeats may lead to missing important biological variants.[1]

2. METHODS
To explore how sequence ambiguity affects read mapping, we have performed several k-mer analyses on MTB genome. We use Mycobacterium Tuberculosis (MTB) H37Rv strain as the reference genome which is 4,411,532 base pairs long.

First, we search for repeated regions (subsequences) in MTB genome using MUMmer 3.0.

Secondly, for finding how complex the MTB genome is and how likely a k-mer can be repeated in multiple places on the genome, we find the number of distinct 25-mers on MTB genome. Then, the number of Hamming distance 1 and 2 neighbours of each distinct k-mer is counted.

Thirdly, in order to investigate the impact of k-mer length on distinctness of the k-mers in MTB, we find the number of distinct k-mers in MTB genome for different k’s.

Finally, to find the distribution of number of places on the reference genome where a read can be mapped with a certain edit distance, we find the Levenstein distance between a random set of k-mers and all k-mers on the reference genome.

3. RESULTS AND DISCUSSION
Firstly, the frequency of repeats with different lengths in MTB genome, shows that, as expected, the number of repeats decreases for larger repeats. However, we can observe some spikes in large repeat lengths which show frequently repeated regions of large size.

Secondly, it can be seen that there are more than a thousand distinct 25-mers that will have about ten repeats in the genome with only one mismatch. This shows how vastly a single nucleotide variation or error can cause ambiguity in read mapping.

Thirdly, we can observe that for k’s greater than 20 the number of distinct k-mers does not change greatly and therefore selecting k = 25 for the k-mer size is a sensible choice.

Finally, a true biological k-mer or read from MTB genome has a higher chance of having a lower edit distance with a k-mer on MTB genome, compared with the edit distance between two random sequence k-mers.

4. CONCLUSION
In general, the MTB genome is fairly stable in terms of complexity for read mapping. However, there are a considerable number of regions in MTB genome which are problematic for read mapping.

5. REFERENCES
Distributed and Cloud Computing
A Responsive Knapsack-based Algorithm for Resource Provisioning and Scheduling of Workflows in Clouds

[Extended Abstract]

Maria A. Rodriguez and Rajkumar Buyya
Cloud Computing and Distributed Systems (CLOUDS) Laboratory
mariaars@student.unimelb.edu.au, rbuyya@unimelb.edu.au

ABSTRACT
Scientific workflows are large-scale, resource intensive applications that can greatly leverage the ease-of-access, affordability, and scalability offered by cloud computing. We propose an adaptive, resource provisioning and scheduling algorithm for scientific workflows deployed in clouds. It was designed to address challenges specific to clouds such as the pay-as-you-go model, the performance variation of resources and the on-demand access to unlimited, heterogeneous virtual machines. It is capable of responding to the dynamics of the cloud infrastructure and is successful in generating efficient solutions that meet a user-defined deadline and minimize the overall cost of the used infrastructure. Our simulation experiments demonstrate that it performs better than other state-of-the-art algorithms.

Categories and Subject Descriptors
C.2.4 [Distributed Systems]: Distributed Applications

Keywords
Cloud, resource provisioning, scheduling, workflow.

1. INTRODUCTION
Clouds offer a scalable infrastructure for the deployment of large-scale scientific workflows. Scheduling algorithms must consider several key characteristics of clouds in order to produce efficient solutions. The first one is the on-demand, elastic resource model, which creates the need for a resource provisioning strategy that works together with the scheduling algorithm. Another feature to consider is the utility-based pricing model used by cloud providers. The cost of using the infrastructure needs to be considered or otherwise, users risk paying prohibitive and unnecessary costs. A third characteristic of clouds is their dynamic state. An example is the variability in CPU performance exhibited by VMs, as well as the network bandwidth degradation and the VM provisioning and deprovisioning delays. This uncertainty needs to be considered when making scheduling decisions.

As a result, we propose the Workflow Responsive resource Provisioning and Scheduling (WRPS) algorithm. Our solution finds a balance between making dynamic decisions to respond to changes in the environment and planning ahead to produce better schedules. Simulation results demonstrate it is scalable in terms of the number of tasks in the workflow, it is robust and responsive to the cloud performance variability and it is capable of generating better quality solutions than static and dynamic state-of-the-art algorithms.

2. THE WRPS ALGORITHM
WRPS has dynamic and static features. Its dynamicity lies in the fact that the scheduling decisions are made at runtime. This allows it to adapt to unexpected delays caused by poor estimates or by environmental changes such as performance variation, network congestion, and VM provisioning delays. The static component expands the ability of the algorithm from making decisions based on a single task to making decisions based on a group of tasks. The purpose is to find a balance between the local knowledge of dynamic algorithms and the global knowledge of static ones. This is achieved by introducing the concept of pipeline and by statically scheduling all of the tasks in the execution queue at once. In this way, WRPS is able to make better optimization decisions and find better quality schedules.

To schedule queued tasks, WRPS groups them based on their type and deadline. Scheduling these bags of tasks is much simpler than scheduling a workflow. There are no dependencies, the tasks are homogenous, and have to finish at the same time. We model the problem of running these tasks before their deadline and with minimum cost as a variation of the unbounded knapsack problem and find an optimal solution using dynamic programming. The same concept is applied to pipelines, they are grouped into bags and scheduled in the same way as bags of tasks are.

We have therefore designed an algorithm which is dynamic to a certain extent in order to manage the unpredictability of cloud environments but that also has a static component that enables it to generate better quality schedules and meet deadlines at lower costs. Moreover, it combines a heuristic-based approach with dynamic programming in order to be able to process large-scale workflows in an efficient and scalable manner.

3. EXPERIMENTS AND RESULTS
The simulation experiments conducted with four well-known workflows show that our solution has an overall better performance than the state-of-the-art algorithms, SCS [1] and SPSS [2]. On average, it succeeds in meeting the deadline constraint in 87.5% of the cases while SCS succeeds in 56.25% and SPSS on 37.5%. These results are inline with what was expected of each algorithm. The static approach (SPSS) is not very efficient in meeting the deadlines whereas the dynamism in WRPS and SCS allows them to accomplish their goal more often. The experiments also demonstrate the efficiency of WRPS in terms of its ability to generate low cost solutions. It outperforms SCS an SPSS as in all of the scenarios except one.

4. REFERENCES
Optimum Benefit Protocol: A Fast Converging, Bandwidth-Efficient Decentralized Similarity Overlay

[Extended Abstract]

Irum Fahim Bukhari syeda.irum@gmail.com Aaron Harwood aharwood@unimelb.edu.au Shanika Karunsekera karus@unimelb.edu.au

ABSTRACT
In this work, we propose a new SBC overlay protocol, OBP, which is based on optimum benefit strategy. It converges rapidly and offers a bandwidth conserving solution that saves 80% communication cost.

Categories and Subject Descriptors
H.3.4 [Information storage and retrieval]: Systems and software—Distributed systems

Keywords
Semantic overlays, gossip protocols, peer-to-peer networks, convergence, bandwidth

1. INTRODUCTION
Due to large volumes of data available on-line, techniques such as document classification and clustering are required for organization, analysis and management of data. Similarity-based Clustering (SBC) is used by many existing systems for information filtering. Decentralized gossip-based overlays offer a simple, robust and scalable solution to SBC clustering. Convergence and communication complexity are the two key areas of concern when SBC is implemented using these overlays. Convergence guarantees accurate clustering and hence, fast convergence is desirable for the applications using these overlays. Communication overhead can become a bottleneck for nodes low in resources. One of the drawbacks of existing SBC protocols is that for similarity distribution with long tail, as power-law distribution, low in-degree nodes take a long time to converge. We refer this as the long tail convergence problem. In this work, we address this problem and propose a new SBC approach, Optimum Benefit Protocol (OBP), that converges more rapidly than existing approaches [1], [2] and reduces the long tail. Our work only sends messages that could benefit the receiver, saving over time about 80% bandwidth. Our protocol obtains at least 90% convergence for a 900 node network, starting in a random configuration, in less than 10 cycles, for all observed experiments. We used synthetic as well as real world distributions from Yahoo, Movielens, and Epinion datasets, for experimentation.

1.1 Experiments
The proposed protocol is analysed using four settings 1. UL-PSS 2. UL-NoPSS 3. Fixed-PSS 4. Fixed-NoPSS. Here, UL is the abbreviation of unlimited view size, and PSS represents peer sampling service. The Figure 1 shows convergence for OBP against all four settings. In Figure 2, OBP is compared with two existing SBC protocols for convergence. The reduction in communication cost for OBP, is shown in Figure 3 as reduction in average message load.

Figure 1: Convergence for OBP settings 1, 2, 3 and 4( Left to right), with Yahoo dataset.

Figure 2: Convergence for OBP, Vicinity, and TMan with Movielens dataset.

Figure 3: OBP average message load (request).

2. REFERENCES
ABSTRACT

Internet and data are growing quickly more than ever and Big Data, which implies huge number of data with velocity, variety, volume and veracity, has brought new challenges in terms of processing and using these data. There are various approaches to process large-scale data sets out there, but graph processing could gain more serious attentions among many of different methods. In this paper, first we explain the importance of graph processing and why it is going to be the dominant processing method of Big Data in the future. Then, we describe the problem that we are going to solve and provide solutions for that.

Categories and Subject Descriptors
C.2 [Computer-Communication Networks]: Distributed Systems; G.2.2 [Graph Theory]: Graph algorithms, Network problems

Keywords
Cloud computing, big data, large graph processing

1. INTRODUCTION

A graph $G = (V, E)$, consists of a set of vertices, $V = \{v_1, v_2, \ldots, v_n\}$ and a set of edges (pairwise relationships), $E = V \times V$. If $(v_u, v_i) \in E$, then $v_i$ and $v_1$ are neighbors. So, $V$ and $E$ are two main characteristics of a graph which most of graph processing frameworks have been implemented based on them. Graphs are widely used in many applications and environments such as social networks, search engines, machine learning, telecommunications, mobile computing, computer networks and datacenters, astronomy, medical research, etc.

As these graphs are growing bigger and bigger, traditional processing approaches are not useful and efficient anymore. So, distributed systems, and cloud computing in particular, can propose seamless approaches for large-scale graph processing.

2. WHY GRAPH PROCESSING?

The reasons that graph processing will be the dominant processing method in the future include: 1) According to an international study that conducted by The National Research Council of the National Academic, graph processing is one of the seven computational giants of massive data analysis, 2) Continuously growing number of graph databases such as Neo4j, InfiniteGraph, MapGraph, OrientDB, etc., which store and analyze data in the form of graphs, shows that many companies and organizations are adopting their systems with this types of databases rather than traditional relational or object oriented databases, 3) According to a recent report from DB-Engine, which is an industry observer, “Graph DBMSs are gaining in popularity faster than any other database category” which shows 300% growth between January 2013 and January 2014.

3. THE PROBLEM AND THE SOLUTION

In this research, we have a large graph with millions or billions of vertices and edges that want to be processed with different algorithms on a public cloud. The algorithms might be convergent or non-convergent. The processing system will be implemented using Amazon cloud infrastructures and platforms. The purpose is to reduce the total execution runtime and the total cost of utilizing cloud facilities. As shown in the above picture, the graph has been stored on Amazon S3. The whole scenario could be as follow: 1) According to a specific initial formula or a random approach, the graph would be partitioned into smaller partitions and a maximum number of a specific type of virtual machines would be selected. 2) Each partition would be assigned to one virtual machine. (The number of virtual machines depends on the number of partitions which depends on the size of each virtual machine’s memory). 3) This is a master-worker model in which one VM would act as the master and other VMs are workers. Master VM is responsible for sending and receiving the reports about processes completions in the workers, synchronizing the VMs, repartitioning the non-processed parts of the graph and so on. 4) Each VM processes its own partition, sends a Finish message to the master and waits until all other VMs complete their works. 5) Before going to the next superstep, on one side according to the number of messages that have been transferred between each pairs of VMs, and on the other side according to the recent changes (updates) of the graph, the master decides to repartition the remaining graph (non-processed parts) based on a novel partitioning algorithm. 6) According to the size of remaining part of the graph, we can terminate a number of VMs. 7) Repeat the steps 4 to 6 until the processing operation converges. 8) For the algorithms that will not converge, such as PageRank, only the steps 4 and 5 are repeated and we don’t terminate VMs. The whole operation will be completed after a user specified number of supersteps.

4. REFERENCES


Auto Scaling within Apache Storm

[Extended Abstract]

Xunyun Liu
Cloud Computing and Distributed Systems Lab
The University of Melbourne, Australia
xunyunl@student.unimelb.edu.au

Rajkumar Buyya
Cloud Computing and Distributed Systems Lab
The University of Melbourne, Australia
rbuyya@unimelb.edu.au

ABSTRACT
Storm is an Apache top-level project which allows performing transformations and analysis on incoming data continuously. As a nature of real-time processing, it is difficult to predict how much computing resources are needed for a given stream processing platform and the trend of the possible varying workload. Unfortunately, the current Storm platform does not provide a capability of automatically scaling out/in to work at an acceptable latency threshold. The goal of this work is to automate the scaling decision within Storm by dynamically probing rational parallelism hints and creating or shutting virtual machines so as to keep the physical resources consistent with the requirement.

Categories and Subject Descriptors
D.4.7 [Organization and Design]: Distributed Systems

Keywords
Apache Storm, auto-scaling, distributed event processing

1. INTRODUCTION
There are two different entities involved in developing and running a stream processing job [1]. The service user designs and implements topologies that suit their need, and then submits those topologies to stream processing service provider who will be responsible for running those topologies on cloud. Users may want to achieve the following goals as a development: a) Be able to deal with possible varying workload without having to change topologies on their own, b) Do not have to specify parallelism hint for each component in advance, since the best specifications of topology under different workloads are hard to speculate and may require a non-trivial tuning process. (In other word, it is the PaaS provider’s responsibility to determine how would the topology be running on their platform). c) Have some basic service level agreement with the provider, the most significant one would be upon end to end latency, which could be calculated by combining with the time that a tuple spends waiting in the input queue and the monitored latency of storm topology.

Similar to a lot of other real-time data processing services, i.e. amazon kinesis and Microsoft Stream Analytics, our service provider provides stream processing as a service and mainly charges users based on the throughput that traverses the topology. Thus they have a strong motivation to minimize the actual resource usage in a long term while keeping SLA as required. Our auto-scaling module is designed to be a handy tool for stream service provider which can be used to adjust platform core numbers (only homogenous machines are considered at first place) and change parallelism hint for components of topology to get the best throughput in that particular resource configurations.

2. DESIGN OF AUTO-SCALING MODULE
In this context, the research questions are when and how to scale the platform and allocate available process unit among the components. Our module is proposed to help service provider to make these decisions by using the famous Monitor, Analyze, Plan, Execute (MAPE) loop to control the system. Since the users have offloaded auto-scaling duty to the service provider, it is necessary to set the auto-scaling module topology-performance agnostic and leave the user out of the auto-scaling process. Our module should remain transparent to the developer and there is no code modification on topology to achieve auto-scaling ability on storm. On the other hand, dynamically adjusting the platform and topology in the learning phase may take times and incur SLA violations, so the service provider should try to shorten the learning process and avoid trial and error process as much as possible.

2.1 System Architecture
Our scaling policy is a combination of naive scaling and pseudo online scaling. Naive scaling can ensure SLA be met during the learning phase with extra short time cost. Since we have no assumption on workload and no knowledge of scaling performance, naive scaling is the best practice we can choose at this stage. At a proper time naive scaling should be replaced by supervised scaling with learned model to gain long term cost saving. Since all the topologies are supposed to be running for a long time, the extra cost of learning process can be considered as negligible. Furthermore, those two methods can benefit from each other. By aggregating data from previous scaling practice, we can have a basic knowledge about the initial state of topology new submitted to the platform and reduce the cost of naive scaling. On the other hand, naive scaling can shed light on which interval the pseudo online learning should be looking at so that the region of search could be minimized.

3. REFERENCES
A Context Sensitive Offloading Scheme for Mobile Cloud Computing Service

[Extended Abstract]

Bowen Zhou
bowenz@student.unimelb.edu.au

Rajkumar Buyya
rbuyya.unimelb.edu.au

Cloud Computing and Distributed Systems (CLOUDS) Laboratory
The University of Melbourne, Australia

ABSTRACT

Mobile cloud computing (MCC) has drawn significant research attention in recent years. In this paper, we propose a prototype MCC offloading system that considers cloud resources such as mobile ad-hoc network, cloudlet and public clouds to provide an adaptive MCC service. Also, we propose a context-aware offloading decision algorithm aiming to provide code offloading decisions at runtime on selecting wireless medium and which potential cloud resources as the offloading location based on the device context.

Categories and Subject Descriptors
C.2.4 [Information Systems Applications]

Keywords
Mobile Cloud Computing; Code Offloading; Context-awareness

1. INTRODUCTION

Mobile cloud computing (MCC) provides services by bringing the abundant resources in cloud computing [1] to the proximity of mobile devices so as to empower the mobile applications performance and conserve the battery life. One of the techniques adopted in mobile cloud computing is code offloading. It identifies the computing intensive code of a mobile program and offloads the task to a cloud service via wireless networks. In the concept of code offloading, cloud resources used for offloading have many different types. First and the most common resource is public cloud computing services like Amazon, Google and Microsoft Azure that provide pay-as-you-go services over the Internet. Secondly, a nearby server named cloudlet [5] is considered as cloud resource with fast network connection as well as powerful processors. Cloudlet serves as a middle layer between mobile devices and public cloud services to reduce the network delay and accelerates the computing. Third, a local mobile device ad-hoc network forming a device cloud [2] is another potential cloud resource, especially when there is no access to the Internet.

Many works have been done in MCC [4] and [3]. They mainly focus on the code partitioning and offloading techniques, assuming a stable network connection and sufficient bandwidth. However, the context of a mobile device, e.g. network conditions and locations, changes continuously as it moves throughout the day. The way of utilizing wireless interfaces can significantly impact the performance of the mobile cloud system and user experience. Moreover, as we described above, there are multiple options of cloud resources that can be selected for code offloading under different conditions. This issue has not been rigorously studied in the literature as many works only target the public cloud service as the resource.

To tackle the issues mentioned above and improve the service performance in mobile cloud computing, we propose a context-aware MCC system that takes the advantages of both nearby cloudlet, local mobile device cloud, and public cloud computing services in the remote to provide an adaptive and seamless mobile code offloading service. The main contributions of our work are as follows:

- First, we present the design of the architecture with nearby mobile cloud, cloudlet and public cloud VMs.
- Second, we provide cost estimation models for cloud VM, local mobile ad-hoc network, and clone VM running on cloudlet and public cloud, and devise an context-aware decision making algorithm taking into consideration the estimation results from the models and the mobile device context to provide offloading policies of where, when and how to offload for the mobile applications.

2. REFERENCES

Neural Network Approach for Universal Dependency Parsing

[Extended Abstract]

Long Duong
Computing and Information System Department
University of Melbourne
lduong@student.unimelb.edu.au

ABSTRACT
Accurate dependency parsing requires large treebanks, which are only available for a few languages. We propose a method that takes advantage of shared structure across languages to build a mature parser using less training data. We propose a model for learning a shared “universal” parser that operates over an inter-lingual continuous representation of language i.e. “universal language”, along with language-specific mapping components. Compared with supervised learning, our methods give a consistent 8-10% improvement across several treebanks in the low-resource simulations.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: Discourse

Keywords
Transfer learning, join training, dependency parsing, resource-poor NLP

1. INTRODUCTION
Figure 1: Dependency tree of the same sentence.

Dependency parsing is the task for understanding the sentence structure. It shows the dependencies between words. For example, the sentence “I like a big meal” has the representations as shown in Figure 1. Dependency parsing is usually used as the core component in the pipeline for Natural Language Processing (NLP), applied successfully for many tasks such as text classification, relation extraction, question answering, statistical machine translation and sentiment analysis. A mature parser normally requires a large treebank for training, however such resources are rarely available and are costly to build. Ideally, we would be able to construct a high quality parser with much less training data, thereby enabling accurate parsing for low-resource languages.

2. METHODS AND RESULTS
In this paper we formalize the dependency parsing task for a low-resource language as a domain adaptation task built upon the work of [1]. A target resource-poor language treebank is treated as in-domain, while a much larger treebank in a high-resource language forms the out-of-domain data. In this way, we can apply well-understood domain adaptation techniques to the dependency parsing task. However, a crucial requirement for domain adaptation is that the in-domain and out-of-domain data have compatible representations. In applying our approach to data from several languages, we must learn such a cross-lingual representation. Here we frame this representation learning as part of a neural network training. The underlying hypothesis for the joint learning is that there are some shared-structures across languages that we can exploit. Compared with the baseline supervised model, our approach give the consistent 8-10% improvement across many languages. More importantly, our approach created a shared cross-lingual representation so that each word is represented as a dense vector (embedded). We expect that words having similar semantic/syntactic properties should have similar representation. Table 1 shows some examples of English-French cross-lingual words embedding. We can see that English and French are highly related and similar to the target word proving the correctness of our model.

<table>
<thead>
<tr>
<th>Words</th>
<th>X-embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>En</td>
<td>Fr</td>
</tr>
<tr>
<td>school</td>
<td>universitaire</td>
</tr>
<tr>
<td>education</td>
<td>université</td>
</tr>
<tr>
<td>student</td>
<td>école</td>
</tr>
<tr>
<td>medicine</td>
<td>scolaire</td>
</tr>
<tr>
<td>participant</td>
<td>school</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initially</th>
<th>Originally</th>
<th>Previously</th>
<th>Actually</th>
<th>Already</th>
</tr>
</thead>
<tbody>
<tr>
<td>originally</td>
<td>réellement</td>
<td>déjà</td>
<td>dernièrement</td>
<td>surroit</td>
</tr>
</tbody>
</table>

Table 1: Example of 5 nearest neighbours with the target English word using our cross-lingual dependency based word embedding.

3. REFERENCES
Bilingual Lexicon Induction from Small Quantities of Phonemic Transcriptions

[Extended Abstract]

Oliver Adams
The University of Melbourne
o.adams@student.unimelb.edu.au

ABSTRACT

We investigate the problem of bilingual lexicon induction from foreign phoneme transcriptions leveraging sentence aligned English transcriptions. Our evaluation of three methods demonstrates that bilingual lexical entries can be learnt with high precision from a corpus of just 10k sentences.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: Machine Translation

Keywords
Lexicon induction, phrase alignment

1. INTRODUCTION

Traditional approaches to language documentation are labour-intensive, requiring much one-on-one time between a field linguist and the mother tongue speakers. Unfortunately, there aren’t enough linguists to document the world’s languages using these approaches before many of the approximately 7,000 languages die out.

Incorporating automatic lexicon induction techniques would speed up the language documentation process. By automatically learning bilingual lexical entries, subsequent data collection may be guided more efficiently since deficiencies in the lexicon highlight aspects of the language inadequately covered in the original audio recordings.

Previous work on bilingual lexicon induction using sentence-aligned corpora has focused primarily on large corpora of written text. However most endangered languages have no established orthography, with audio data and phonemic transcriptions being easier to acquire. Bilingual lexicon induction applied to phonemically transcribed audio introduces problems, including the lack of word segmentation and small quantities data.

We investigate the problem of automatically extracting bilingual lexicons from phonemic transcriptions using data that takes the form of unsegmented phonemes sentence aligned with English words. Existing methods based on word alignment face issues due to a poor relationship between individual source phonemes and target words. We mitigate this issue by applying two new approaches and comparing them to a baseline word alignment approach with GIZA++ [3].

The first approach (UWS GIZA++) performs unsupervised word segmentation using a Bayesian Pitman-Yor language model as implemented in the tool pgibbs [1] before alignment. This is more effective since it breaks foreign phoneme sequences into coarser tokens that translate better to English.

The second performs joint word segmentation and alignment using the substring alignment model of [2], as implemented in pialign. The advantage of this approach over the previous two methods is that the segmentation on the phoneme side can be informed by the English.

We run experiments to assess the induced lexicons’ precisions at k entries by manually annotating the lexicons. Results demonstrate that hundreds of bilingual lexical entries can be learnt with precisions superior to those achieved with a traditional alignment approach, offering promise of the technique’s applicability in a language documentation context.

2. ACKNOWLEDGEMENTS

Graham Neubig, Trevor Cohn and Steven Bird.

3. REFERENCES


1http://github.com/neubig/pialign
Let’s Shoot the Breeze!
Flexible Language Independent Multiword Expression Analysis

[Extended Abstract]

Bahar Salehi
NICTA Victoria Research Laboratory
Department of Computing and Information Systems,
The University of Melbourne
bsalehi@student.unimelb.edu.au

ABSTRACT
In this paper, we discuss language independent and highly effective methods for measuring the compositionality of multiword expressions (MWEs).

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]

Keywords
Multiword expressions, String similarity, Distributional similarity, Word embeddings, Deep learning

1. INTRODUCTION
A multiword expression (MWE) is defined as any combination of words with lexical, syntactic or semantic idiosyncrasy, in that the properties of the MWE are not predictable from the component words. For example, with shoot the breeze, we have semantic idiosyncrasy, as the meaning of “to chat” in usages such as It was good to shoot the breeze with you cannot be predicted from the meanings of the component words shoot and breeze.

Among the interesting features of MWEs, their semantic idiosyncrasy (or compositionality) has been of particular interest to NLP researchers. An MWE is fully compositional if its meaning is predictable from its component words, and it is non-compositional (or idiomatic) if not. For example, stand up “rise to one’s feet” is compositional, because its meaning is clear from the meaning of the components stand and up. However, the meaning of strike up “to start playing” is largely unpredictable from the component words strike and up.

In this study, I consider compositionality to be graded, and aim to predict the degree of compositionality. For example, in one of the datasets, climate change is judged to be 99% compositional, while silver screen is 48% compositional and ivory tower is 9% compositional.

No existing resources, such as dictionaries, include explicit information about the precise level of compositionality of MWEs. Most recent work on predicting the compositionality of MWEs can be divided into two categories: general-purpose and language/construction-specific. The bulk of work on compositionality has been language/construction-specific, using specific properties of the MWE in that language to predict compositionality.

In this paper we briefly introduce our three language independent approaches:

String similarity approach.
We proposed a string similarity approach in [1], which is general-purpose and does not require any language specific features. Our results show that this approach performs better than distributional similarity for morphologically rich German noun compounds [2], and hard-to-identify English verb particle constructions such as put on.

Multilingual approach.
Our further studies [2] shows that by combining distributional similarity scores from multiple languages, we can improve monolingual distributional similarity approach. Note that in this approach, unlike the previous studies, we did not use any parallel corpora. Our string similarity approach [1] also shows that integrating information from multiple languages improves compositionality prediction.

Deep learning approach.
Our recent paper [3], proposes the first attempt to bring together the work on word embedding-style distributional analysis (based on deep learning) with compositionality prediction of MWEs.

2. ACKNOWLEDGMENTS
I would like to thank my supervisors Prof. Timothy Baldwin and Dr. Paul Cook for their support and guidance.

3. REFERENCES

The example is taken from http://www.thefreedictionary.com
Human Computer Interaction and Design
Technology at Mealtime: Beyond the ‘Ordinary’

[Extended Abstract]

Hasan Shahid Ferdous
Dept. of CIS, University of Melbourne, Australia
hasan.ferdous@unimelb.edu.au

ABSTRACT

In this research, we investigate the everyday interactions of familial uses of technology around mealtimes and explore how family members configure the dinner space and the technologies within it. We seek to understand how technologies are used and negotiated amongst family members and the influence of technologies on the content and context of their interactions. We aim to pay special attention to understand how our everyday technologies support our regular mealtimes as well as special occasions.

Categories and Subject Descriptors
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Keywords
Commensality; Communication Technologies; Food.

1. INTRODUCTION

Family meals have always been a source of social interaction, cultural heritage, enjoyment, and celebration. While the social foundations of commensality have long been an important concern within sociological and anthropological fields, relatively few works in HCI discussed the need for food-related research to go beyond efficiency, to consider the experience, affect, and desire of eating and sharing time together. In their seminal work [2], Grimes and Harper discussed the creativity, endowment, relaxation, and nostalgia found in the togetherness of family meals. Throughout this thesis, we have taken this celebratory approach to explore and understand the current technological practices around family mealtime to better inform us regarding future design.

2. RELATED WORKS

Hupfeld et al. [3] provided a detailed account of the everyday practices associated with domestic food consumptions and how it relates to the ecology of mealtime artifacts and spaces - both technological and otherwise. Barkhuus & Brown [1] noted that the television viewing is not just a simple form of distraction or removal from the social as it is often portrayed. Rather, it is an activity that is integrated into the broader social practices and arrangement of the household. We use these works as a springboard to understand the ways in which a broader set of everyday technologies become implicated in the social configuration of everyday commensality practices and family relations at mealtimes. Rather than making strong moral arguments with respect to the position of technology within mealtimes, we look to the ways that such technologies contribute or detract from any idealized notions of family order in these settings.

3. FINDINGS FROM STUDY ONE

We base our analysis on the Heideggerian’s terms ‘ready-to-hand’ and ‘present-at-hand’. We identified technologies as ‘ready-to-hand’ when they are immediately available and their usefulness is clearly apparent within the dining context. Technologies that are ‘present-at-hand’ are still available but somewhat removed from the dining activity. Ready-to-hand technologies are available (socially and cognitively) for immediate use; present-at-hand technologies require a little more effort to discern their whereabouts and role in the social milieu.

We observed four patterns of familial arrangement around the furniture and available technologies during mealtimes - i) technologies orientate to people, ii) people orientate to technology, iii) hidden technologies, iv) displaced technologies. Technology often served as a conversational resource during mealtimes - it was, in itself, a topic of conversation. What was also of further interest was that certain technologies were used as a resource for setting certain scenes and creating ambience. Such creative practices were not necessarily about the more everyday and routine but became a way of making special certain occasions. Finally, distraction was in itself not always at odds with the social conduct of the family mealtime, as we saw in the strategic efforts to encourage child satiety or maintain family harmony.

4. CONCLUSION

We have explored current practices around the personal and shared devices and analyzed how they support familial conversation, provide relaxation, and enable celebration, albeit not without occasional tension. In our subsequent studies, we aim to explore the possibilities enabled by sharing the content from personal devices during family mealtimes. We are now investigating how such system can augment the social interactions, mediate usage, and/or bring tension and constraints among the family members.

5. REFERENCES

Finding Time for Tabletop: Boardgame Play and Parenting

[Extended Abstract]

Melissa J. Rogerson
Microsoft Research Centre for Social Natural User Interfaces & Interaction Design Lab, The University of Melbourne
mellissa.rogerson@unimelb.edu.au

Martin Gibbs
Microsoft Research Centre for Social Natural User Interfaces & Interaction Design Lab, The University of Melbourne
martin.gibbs@unimelb.edu.au

ABSTRACT
In this paper, we describe the ways in which serious ‘hobby’ boardgame players continue to make time for their hobby after becoming parents. We identify key strategies that they use, including reservation of dedicated times and locations for play as well as inviting their children to participate in their hobby. Further, we identify digital devices as a key tool that enables continued engagement with the pastime of boardgames, both as a player and through the extended sociality that surrounds play.

Categories and Subject Descriptors
H.5.2 [User Interfaces]: Miscellaneous; K.8.0. [General]: Games; H.1.2 [User/Machine Systems]: Human factors

Keywords
Boardgames, play, parenthood, serious leisure.

1. INTRODUCTION
Both digital and tabletop games attract committed aficionados – passionate insiders or serious leisure practitioners who devote seemingly limitless time and energy to their hobby [4]. We describe these hobby gamers as serious gamers, who place value on the pastime of gaming [1] and on “active participation within the hobbyist culture, the acquisition and accumulation of games, and a degree of evangelism for the hobby.” [5].

Themes of work, relationships, home ownership and family responsibilities all influence adult players’ engagement in both analog and digital games. Although for some, these present a barrier to play or at least to playing certain types of game [see 3], for serious boardgame players these pressures seem more to dictate the choice of game or timing and frequency of sessions than to drive them away from the hobby. Although gamers’ time is restricted, their desire to pursue the hobby through active engagement in its culture remains strong.

2. FINDINGS
In this paper we present findings from a study of serious boardgamers. These are people who approach board gaming as a hobby and as a form of serious leisure (Stebbins, 2012). In particular, we examine how adults negotiate time and opportunity to continue their serious pursuit of board games while also managing their other commitments, particularly as parents. Using data derived from interviews and web forum posts, we investigate the ways in which parents’ tabletop play is curtailed as well as the strategies that they use to carve out time for boardgaming.

Common approaches that are used by many parents include reframing games as an opportunity for cross-generational play. Boardgamers embrace the opportunity to evangelize their hobby, and the idea of “built-in” gaming partners is appealing; parents – especially fathers – often choose to share their hobbies with their children [2]. Others negotiate time and location (at or away from home) of gaming sessions to suit their family’s situation.

But whilst these strategies focus on finding or creating opportunities for face-to-face play, digital devices also offer an outlet for the serious gamer even when restricted to the home. Opponents need not be in the same place as the gamer; asynchronous games ensure that the opponent need not even be in the same timezone. Players can engage with digitized games and opponents within their home, even whilst spending time with their children or spouse. Despite a sense that digitized games are somehow less real, we have found evidence of broad adoption of digitized boardgames by parents, particularly of younger children, as a means of maintaining a connection with their hobby.

3. CONCLUSION
Serious adult boardgamers demonstrate the desire to continue to engage with their hobby, and have developed a host of strategies to continue their engagement as the responsibilities of parenthood intervene. Our research identifies strategies used for ensuring face-to-face play, and further shows that digitized boardgames are used as a strategy for maintaining active participation in the boardgame community as well as with the hobby through changing life phases.

4. ACKNOWLEDGMENTS
This research was supported in part by the Microsoft Centre for Social Natural User Interfaces. The authors gratefully acknowledge the support of the Centre.

5. REFERENCES
Clinicians’ perceptions of computerised support for mental health

[Extended Abstract]

Fernando Estrada
Dep of CIS, University of Melbourne
festrada@student.unimelb.edu.au

ABSTRACT

Substantial research reports that computerised support for mental health can assist individuals suffering from a mental health disorder in their recovery. However, high levels of attrition from such applications has been reported and an in-depth understanding of the reasons users cease use is yet to be developed. However, research into applications for mental health report that users are more inclined to use, and continue using an application if they are encouraged by mental health professionals.

Therefore, the aim of this research study was to identify and understand clinicians’ assumptions, beliefs, understanding of and experiences with such applications as a means to understand the role that clinicians play in users’ attrition from such applications. At this stage in this study we are analysing data obtained from interviews with clinicians using a thematic analysis approach to identify and explain emerging themes and subthemes. The overall aim of this study is to provide recommendations on the better design and evaluation of computerised support for mental health.

Categories and Subject Descriptors

H.4.0 [Information Systems Applications]: General

Keywords

Mental Health, clinicians, experience, use, engagement, technology

1. INTRODUCTION

The World Health Organization reports that mental health disorders are one the leading causes of disability. This is particularly concerning as it has been estimated that 25% of the world’s population may be experiencing mental health disorders and resources for face-to-face consultations with mental health professionals (“clinicians”) are diminishing [1]. In this context, some people with mental health disorders (“users”) are turning to the use of computerized support (“applications” available on desktop, mobiles and tablet devices) for mental health as a source of support. In Australia, there appears to be an increasing willingness to access this kind of support. However, high levels of attrition from such applications has been reported and an in-depth understanding of the reasons users cease use is yet to be developed.

2. CLINICIANS’ PERCEPTIONS

Substantial research reports that users are more inclined to use, and continue using such applications when encouraged by clinicians [2]. Doherty et. al. [2] reports that clinicians are more willing to recommend the use of such applications if they see the value it provides to users. Additionally, research into applications for mental health has identified a need to ensure that such applications are endorsed by mental health institutions, align with clinicians’ psychological approach and support clinicians’ relationships with users. However there is still a paucity of knowledge in relation to clinicians’ assumptions, belief, understanding of and experiences with such applications. This is significant as clinicians’ perceptions may impact users’ use of such support.

3. RESEARCH DESIGN

We obtained ethics approval from the Melbourne Clinic to conduct this study (project approval #246). The aim was to gain an understanding of clinicians’ assumptions, beliefs and views in relation to applications for mental health as a means to understand the role that clinicians play in users’ attrition from such applications. We conducted 16 interviews with clinicians who all provide face-to-face consultations to patients of the clinic. The following questions are indicative of the questions that we asked them: Do your clients use mental health applications? (If so, which ones?) Have your patients discussed mental health applications with you? Do they get value from using these applications? Have they asked you for your opinion on these applications? What are your feelings about your clients’ use of these applications? Do you feel you are sufficiently educated about mental health applications to discuss them with your clients? Where do or would you go to find out information about applications? Do you ever recommend to clients that they use applications for mental health? If so, which ones? What is your opinion of these applications? Where do these applications fit within your client’s treatment? What do you consider are the benefits of using such applications? If not, why do you not recommend the use of such applications?

3.1 Research Analysis

We used a thematic analysis approach to analyse the data obtained from the interviews. To date, we have identified the following seven themes: outcomes, collaboration, models, engagement, security, privacy and protocols that have emerged from the interviews. The overall aim of this study is to provide recommendations on the better design and evaluation of applications for mental health.

4. REFERENCES


Understanding the Current Practices of Teleconsultation

[Extended Abstract]

Deepti Aggarwal
Interaction Design Lab
University of Melbourne, Australia
daggarwal@student.unimelb.edu.au

Abstract
This research explores the situated use of technology by patient, clinician and other supporters in the context of teleconsultation, and highlights the issues and challenges that limit the patient-doctor interactions.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: UI

Keywords
Teleconsultation; Telehealth; Ethnography field study

1. INTRODUCTION
Teleconsultation is the video mediated medical consultation between a patient and clinician for the purpose of diagnostic and therapeutic advice [5]. Although teleconsultation has been around for over a decade or two, the academic understanding of its practices has been relatively limited. There have been preliminary studies on teleconsultation that aims to understand mainly its technical feasibility and cost effectiveness [1,3]. However, lesser attention is paid on understanding the operational practices and exchange of information during consultation. Taking inspiration from face-to-face consultations where physical interactions play a crucial role in understanding the health issue of the patient, I argue that verbal communication during teleconsultation may not be sufficient in supporting medical diagnosis during teleconsultation. This research, therefore, investigates the current practices of teleconsultation to understand how technology support or inhibit the essential interactions between a doctor and patient during a consultation.

2. APPROACH
Major goal of this research is to explore how the available teleconsultation systems support doctor-patient communication over video. In this regard, I conducted an exploratory study in collaboration with two major medical organizations operating in Melbourne: Royal Children’s Hospital (RCH), and Inner East Melbourne Medicare Local (IEMML). Study participants included patients, clinicians and other supporters such as caregivers, and clinician’s assistants. I have utilized ethnographic approach [4] to understand the study context where I observed both face-to-face and teleconsultation sessions and conducted semi-structured interviews with the participants. Observations of teleconsultation and face-to-face consultation sessions offered me an understanding of how the medical practices differ when moved from a physical space (face-to-face) to a video communication. Additionally, interviews offered me insights about the subjective experience of participants with teleconsultation systems. Field notes were taken during the sessions, while interviews were audio recorded for later analysis. I have utilized Grounded Theory approach [2] to interactively collect and analyze the study data.

3. FINDINGS
I have observed 10 sessions: 7 teleconsultation and 3 face-to-face consultations and 12 semi-structured interviews with the participants. Findings of the study are presented under the following five themes:

Understanding of the infrastructure: How are the teleconsultation sessions organized currently? For what purposes, teleconsultations are used? What is the physical arrangement of people and technology during teleconsultation?

Understanding of the actors: Who are involved in the teleconsultations? What are their respective roles? What are their levels of competency? What issues do they face and how do they try to resolve them, while having teleconsultation?

Understanding of the interactions: What types of interactions happen among people during teleconsultation? How do people interact with the underlying technology during teleconsultation?

Understanding of the technology: What sorts of technologies are currently used? How the technology is changed over a period of time? How is technology decided to consult a given medical issue over teleconsultation?

Contrasting teleconsultation with face-to-face consultation: If and how each of the above (infrastructure, technology, actors, and interactions) themes differs in teleconsultations when compared with face-to-face consultations?

The study has also unveiled several opportunities to design better interactive systems to support doctor-patient interaction in teleconsultation. Understanding gained from this study will be utilized to design a tangible interface to enhance communication and interaction during teleconsultation.

4. REFERENCES
Machine Learning and Data Mining
Tour Recommendation on Location-based Social Networks

[Extended Abstract]

Kwan Hui Lim
The University of Melbourne and NICTA
limk2@student.unimelb.edu.au

Christopher Leckie
The University of Melbourne and NICTA
caleckie@unimelb.edu.au

Jeffrey Chan
RMIT University, Australia
jeffrey.chan@rmit.edu.au

Shanika Karunasekera
The University of Melbourne
karus@unimelb.edu.au

ABSTRACT
We propose two tour recommendation algorithms based on the Orienteering problem, and implement a tour recommendation framework that uses geo-tagged photos and Wikipedia to determine the travel histories and interests of tourists.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications - Data mining; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithms, Experimentation, Measurement

Keywords
Tour Recommendation, Travel Itinerary, User Interests, Orienteering Problem, Flickr, Wikipedia, Social Networks

1. INTRODUCTION
For a visitor in a foreign city, it is a challenging task to plan a tour or itinerary that is customized to his/her interest preferences. Despite the availability of online resources, there are challenges in planning a customized tour, namely: (i) many travel guides recommend popular Points of Interest (POI) that are not personalized to the tourist’s interest preferences; (ii) even with a list of POIs, it is a tedious task to construct an itinerary based on those POIs with the considerations of travelling time and specific starting/ending points. Our work addresses these challenges by proposing two algorithms based on variants of the Orienteering problem and implementing a tour recommendation framework (Fig. 1) that utilizes geo-tagged photos and Wikipedia.

2. PROPOSED APPROACH
We model the tour recommendation problem as a variant of the Orienteering problem [3], with a time/distance budget B, starting POI p1 and destination POI pN. Our goal is to recommend a tour T = (p1, ..., pN) that maximizes POI popularity and tourist interest, while staying within the budget B. Specifically, we proposed the following algorithms:

- **PersTour Algorithm** [2]: For recommending personalized tours with POIs and visit duration based on POI popularity and time-based user interests.
- **TourRecInt Algorithm** [1]: For recommending customized tours with a mandatory POI category based on user/tourist interests.

3. EXPERIMENTS AND RESULTS
We evaluate the effectiveness of PersTour and TourRecInt against various greedy-based baselines on a Flickr dataset of multiple cities. Our experimental results show that PersTour and TourRecInt out-perform the baselines in terms of various evaluation metrics, namely: tour popularity, tour interest, POI count, precision, recall, F1-score, and RMSE of visit duration.

For more information on our proposed algorithms and experimental results, please refer to [1] and [2].

Acknowledgments. This work was supported by NICTA.

4. REFERENCES
Privacy-Preserving Location Mining of Tweets

[Extended Abstract]

Shuo Wang
Department of Computing and Information Systems
University of Melbourne
shuow4@student.unimelb.edu.au

Richard O. Sinnott
Department of Computing and Information Systems
University of Melbourne
rsinnott@unimelb.edu.au

ABSTRACT
With the global adoption of smart mobile devices equipped with localization capabilities and broad popularity of microblogging facilities like Twitter, the need for personal privacy has never been greater. This is especially so with computational and data processing infrastructures such as Clouds that support big data analysis. Differential privacy of geospatially tagged data such as tweets can potentially ensure that degrees of location privacy can be preserved while allowing location-based pattern mining. We propose a differential location pattern mining approach considering both privacy and precision of geo-located tweets clustered according to Geo-Locations of Interest (GLI).

Categories and Subject Descriptors
K.6 [Information management]: Security & protection

Keywords
Pattern mining, location privacy, Tweets.

1. INTRODUCTION
Twitter has become a global phenomenon with over 400 million tweets made daily. Many users are unaware that often the geolocation of the tweet is also recorded, i.e. where they actually tweeted from and at what time they tweeted. Location-based services are increasingly popular and not restricted solely to Twitter. The availability of major computational resources such as Clouds and big data processing algorithms such as Map Reduce and Elastic Search now allow mining and analysis of data at an unprecedented scale. Given this, it is meaningful to explore behavioral analysis and pattern mining of location data and identify ways to obfuscate this sensitive information, especially as it could be used for malicious purposes. Recently, differential privacy [1] has been widely used for the protection of location-based data. It was shown that location privacy could be preserved by adding moderate degrees of noise based on an appropriate degree of required location obfuscation, while supporting degrees of service for other location-based services. The advantage of differential privacy for location privacy is that it allows to protect individual location information whilst still allowing the data to be used for analysis and/or mining. Solutions that can limit the dangers of leaking location privacy would encourage more users to share their location information. A large amount of meaningful work with social utility could be carried out with improved aggregate geospatially-coded Twitter data, e.g. pandemics and natural disasters often rely on social media and being able to undertake pattern mining to extract knowledge such as GLI of tweets with “safe” degrees of privacy preservation and potentially their followers.

2. DATA AND METHODS
The software architecture used to support the explorations of location privacy of Twitter data supports data collection through the Twitter Search API and Streaming API, data preparation, and the associated methods required to perform ε-differential privacy and GLI pattern mining and associated analysis. Methods used to generate a differential privacy-driven database from raw geolocated Twitter data is achieved in two steps. The first step is to decompose spatial location regions by optimal quad-trees with differential privacy mechanisms. Following this, clustering of intersecting areas to find GLIs with perturbed outputs is undertaken to support differential privacy for locations.

Differential Privacy Spatial Decomposition. The classical solution to ensure differential privacy for spatial points datasets is to decompose the spatial space, and then publish statistics on the points within each region in a differential privacy-preserving way. Users can get obfuscated knowledge of locations by intersecting the query regions of the split areas. The method to build differential privacy spatial decomposition can be divided into adding noise to counts and index structures satisfying differential privacy. The purpose of spatial decomposition is to divide a global task into several local subtasks.

Extracting the Differential Privacy of GLIs. An improved DBSCAN [2] is used to extract GLIs with differential privacy guarantees. The output of this algorithm is the set of privacy preserved GLIs given as the region centroids. The improved DBSCAN has further improvements to DBSCAN that calls DBSCAN with different distance thresholds ε and density threshold MinPts, and returns the result when the number of clusters is appropriate.

3. EVALUATE METRICS
The evaluation metrics used to measure the applicability of the approach described focuses on utility and privacy features of the differential privacy location pattern mining method to discover GLIs. These evaluation metrics contain several aspects: the inferred number of actual GLIs; the Euclidean distance between actual GLIs and privacy-enabled GLIs; the count difference of points in the intersection of real regions and privacy preserving regions, as well as the number of similar neighborhoods surrounding real GLIs and location privacy enabled GLIs.

4. REFERENCES
Generalized Information Theoretic Cluster Validity Indices for Soft Clusterings

[Extended Abstract]

Yang Lei
Department of Computing and Information Systems
University of Melbourne
yalei@student.unimelb.edu.au

ABSTRACT
There have been a large number of external validity indices proposed for cluster validity. One such class of cluster comparison indices is the information theoretic measures, due to their strong mathematical foundation and their ability to detect non-linear relationships. However, they are devised for evaluating crisp (hard) partitions. In this work, we generalize eight information theoretic crisp indices to soft clusterings, so that they can be used with partitions of any type (i.e., crisp or soft, with soft including fuzzy, probabilistic and possibilistic cases).

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Clustering

Keywords
Cluster Validity; Soft Clusterings; Information Theoretic

1. INTRODUCTION
Clustering is one of the most important unsupervised techniques. It aims to divide the data objects into several groups, so that the objects in the same group are similar whereas the objects in different groups are dissimilar. Clustering validation, which evaluates the goodness of a clustering, is a challenging task. There have been a large number of clustering validity measures proposed, which can be generally classified into two categories, internal clustering validation and external clustering validation [2]. In this work, we focus on the external validation measures.

External validation measures are also called similarity measures, used for comparing the similarity between a pair of clusterings. We show an example in Figure 1, explaining how do we measure the goodness of a clustering based on external validation measures. Given a ground truth clustering shown in Figure 1a, how do we choose the best (the most similar one to the ground truth clustering) clustering from clustering 1 to 3 (Figures 1b to 1d)? We can use the similarity measures (external validity measures) to quantify the similarity between each of the three clusterings and the ground truth clustering, and choose the one with the highest similarity. In this case, clustering 2 is the most similar one.

Most external validity indices compare two crisp partitions [2]. However, partitions can also be soft, i.e., fuzzy, probabilistic or possibilistic partitions [1]. Soft partitions are usually converted to crisp partitions. Then they are evaluated by employing the crisp external validity indices. However, this kind of conversion may cause loss of information. Information theoretic measures form a fundamental class of measures for comparing pairs of crisp partitions. They have drawn considerable attention in recent years [3]. However, they are designed for comparing crisp clusterings and cannot be used to compare soft ones.

Therefore, in this work, we aim at generalizing eight information theoretic crisp cluster validity indices to the soft case. It has been demonstrated via experimental evaluation that the generalized information theoretic indices can be useful for comparing the similarity between pairs of soft clusterings.

2. REFERENCES
ABSTRACT

Modularity is a well-known quality function for community detection in complex networks. However, modularity’s limited view on vertex similarity leads to limits in its performance. To overcome these limitations, we propose a generalized modularity measure called GM which has a more sophisticated interpretation of vertex similarity. Experiments on different synthetic and real data sets, demonstrate GM performs strongly in comparison to several existing approaches, particularly for small-world networks.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information search and retrieval-Community detection, Graph Clustering

Keywords

Community detection, Modularity, Vertex similarity

1. OVERVIEW

Among community detection criteria, modularity [1] is one of the most important ones. It measures how vertices in a community share more edges than what would be expected in a randomized network.

Modularity has a restricted interpretation of vertex similarity which is limited to just considering shared edges among vertices. Figure 1 illustrates this problem using hand-made examples. Degrees of the bold vertices are the same in both cases and they do not share any edges together. As the total number of network edges are also the same in both cases, the modularity value between the two bold vertices is equal while one can clearly see that the structures are very different.

Figure 1: Examples for limited vertex similarity interpretation of Modularity

We propose a new measure of community detection called generalized modularity (GM) which extends modularity’s assumption about similar vertices. We believe that sharing more neighbours than what is expected (in a randomized network) also expresses similarity of vertices and shows how likely it is for the pair to lie in the same community. Therefore, in addition to common edges, GM takes into account common neighbours and longer paths between vertices and compares the number of these paths to a randomly distributed network to achieve a more comprehensive interpretation of vertex similarity.

To illustrate how well generalized modularity can reveal the underlying community structures, we use visualization. Figure 2 presents VAT images of modularity and GM for some real and synthetic benchmarks. In this Figure, modularity’s VAT image does not reveal the community structure of the data sets while dark blocks in GM’s VAT image effectively distinguish community structures.

In future, improving GM to overcome other limitations of modularity can be a clear direction.

2. REFERENCES

How Private Can I Be Among Public Users?
[Extended Abstract]

Elham Naghizade, James Bailey, Lars Kulik, Egemen Tanin
Computing and Information Systems Department
The University of Melbourne
Melbourne, Australia
{enaghi,baileyj,lkulik,etanin}@unimelb.edu.au

ABSTRACT

People are increasingly volunteering personal data. Services based on this data rely on a high number of participants and high data quality. Personal data is often seen as private and users are more likely to provide such data if they can choose its granularity, e.g., instead of an exact value, they may provide a range. Focusing on spatial crowdsourced data, this work aims to determine whether the common method of coarsening location data of privacy-conscious individuals is an effective approach if fine-grained location data has also been submitted by privacy-apathetic users.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications

Keywords
Privacy, Crowdsourced Location Data, Matrix Factorization.

1. PROBLEM STATEMENT AND OUR APPROACH

The uptake in sensor-enabled smartphones and wearable devices enables individuals to monitor not only their environment, but also themselves on a 24/7 basis. Individuals have different perceptions of privacy, which has a direct impact on their data volunteering behaviour. Various studies have stressed the importance of accommodating users’ privacy preferences both as a means of encouraging them to share their data and preserving their privacy [1].

In such scenario, privacy-apathetic users may contribute detailed data, while privacy-conscious users may prefer to provide a range instead of a precise value or a cloaked region instead of an actual position. Our key question is to what extent volunteering coarse-grained data can actually guarantee the desired level of privacy for privacy-conscious users.

We focus on location data. Given a set of users who contribute their daily commute patterns with different resolutions respective to their privacy preference, we investigate if it is possible to use the more fine-grained trips of privacy-apathetic users in conjunction with the coarse-grained trips of the privacy-conscious users to refine the contributed data of the latter. Despite many efforts in the literature to address the privacy issues of sharing detailed location data, the privacy implications of having a multi-granular dataset has not been investigated.

We deployed a grid structure [2] to store location data at different granularity levels (Figure ??) since i) it provides a flexible as well as comprehensible means of facilitating users’ preference specification ii) it is independent of the original trajectory and does not reflect any specific property of the data. We store the location data as a sequence of grid cells.

The adversary may extract multiple versions from the multi-granular grid-based dataset through generalization: A complete dataset (Figure ??) that contains all users’ location data at the lowest granularity, i.e., coarsest resolution, and several incomplete datasets (Figure ??) with more fine-grained location data and in which the location data of privacy-conscious users for that specific level is not known.

We propose an matrix factorization (MF) based approach that uses the data available for both granularities for privacy-apathetic users: it learns a transition matrix that maps users’ coarse-grained locations to their fine-grained locations. This transition matrix is then applied to the coarse-grained data of privacy-conscious users to predict their fine-grained location. This makes our method a supervised approach contrary to the unsupervised classical MF. Our inference attack manages to refine the footprint of privacy-conscious users with a high precision and recall.

2. REFERENCES


Modelling Pedestrian Activities in a Dynamic Environment

[Extended Abstract]

Minh Tuan Doan  
National ICT Australia,  
Department of Computing and Information Science  
The University of Melbourne, Australia  
mdoan@student.unimelb.edu.au

Sutharshan Rajasegarar  
National ICT Australia,  
Department of Computing and Information Science  
The University of Melbourne, Australia  
sraja@unimelb.edu.au

Christopher Leckie  
National ICT Australia,  
Department of Computing and Information Science  
The University of Melbourne, Australia  
caleckie@unimelb.edu.au

ABSTRACT
We analyse the normal behaviour of pedestrians over the CBD area of the City of Melbourne, Australia. We addressed three main challenges in this problem: (1) the data is periodic, (2) the system switches intermittently between different states when the data is considered for a long period of time and (3) the data is high dimensional and multiple locations need to be analysed simultaneously. Our approaches demonstrate their ability to correctly identify the anomalous behaviours in the pedestrian movement data.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics

Keywords
Application, Anomaly Detection

1. INTRODUCTION
The City of Melbourne has deployed several pedestrian count sensors throughout the Melbourne CBD. These sensors have been collecting hourly pedestrian counts over 28 locations around the central business district of the city since 2009 [1]. However, no analysis has been performed on this data. We perform a systematic analysis on this data to reveal any interesting behaviour about the pedestrian movement pattern in the city. We modelled the normal behaviours of pedestrian activities in terms of distributions of pedestrians at different locations in the city. We then used the model to detect any anomalous pedestrian distributions. The fact that the input data is multi-dimensional and collected over an extended period raises many research challenges that need to be resolved for an effective model. Our key contributions include: (1) Developing and employing a cluster-based anomaly detection algorithm that copes with systems that switch intermittently between states. The evaluation reveals that the accuracy of this approach outweighs its complexity when compared to a single pass clustering technique. (2) Developing and evaluating a technique based on frequent itemset mining to model the normal behaviour and detect anomalies from high dimensional pedestrian data.

2. CHALLENGES AND APPROACHES
First, the data is periodic on daily, weekly, monthly cycles and the patterns tend to repeat accordingly. Hence merely treating the data as the sequence of observations might potentially hide any interesting patterns in there. Second, a long period of collected data is used for the modelling and this model of normal distributions, which can be affected by multiple external factors such as weather or seasonal holidays, may vary throughout the year. Single pass clustering techniques only provide a static model of the system and do not consider how it changes over time. In order to address the first two challenges, we employed an Ensemble Switching Model [3], which is a clustering model that can cope with systems that switch intermittently between states. By evenly dividing the set of pedestrians into small windows, the algorithm identifies the different states in the data and selects only the relevant states as the basis for anomaly detection.

Third, in order to describe the distribution of pedestrians over a large area, we need to consider observations at multiple locations in the city, resulting in high dimensional input data. Our initial approach which uses clustering algorithms has yielded ineffective results in this case. Hence, we developed an approach that employs frequent itemset mining to construct the model in terms of correlations of pedestrian distributions across multiple locations in the city.

3. RESULT
By analysing only two locations (dimensions), we evaluated the Ensemble Switching Model in comparison with the static clustering technique HyCARCE [2] and then cross referenced with known events to validate the accuracy. It was observed that this dynamic model produces better results and is more immune to false positives and false negatives.

Moreover, we also evaluated the frequent itemset-based approach by analysing ten different locations around the CBD (ten dimensional data). We observed high accuracy and recall rate using this approach.

4. ACKNOWLEDGMENTS
The authors would like to thank the support from National ICT Australia (NICTA).

5. REFERENCES
Making Metric Learning Feasible for Large Datasets

[Extended Abstract]

Zay Maung Maung Aye, Kotagiri Ramamohanarao and Benjamin Rubinstein
Department of Computing and Information Systems
The University of Melbourne
zaye@student.unimelb.edu.au, {kotagiri, benjamin.rubinstein}@unimelb.edu.au

ABSTRACT

Many machine learning and pattern recognition methods rely heavily on good distance metrics to achieve competitive performance. While distance metrics can be learned, the computational expense of doing so is infeasible on large datasets. In this paper, we propose an efficient and effective approach for selecting the training dataset using locality sensitive hashing (LSH) with discriminative information and its variant K-Means clustering inside LSH buckets for accelerating metric learning. Our generic approach can accelerate state-of-the-art metric learning while achieving competitive classification accuracy, expanding feasibility for large data sets by 1-2 orders of magnitude.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning; I.5.4 [Pattern recognition]: Applications; I.5.3 [Pattern recognition]: Clustering; I.5.2 [Pattern recognition]: Design Methodology—Feature evaluation and selection

Keywords
Metric Learning, Locality Sensitive Hashing, Scalable Learning

1. INTRODUCTION

Distance metric learning is an important topic in machine learning. Performance of many machine learning and pattern recognition methods depends significantly on the similarity or distance metric used. The distance metric is problem-specific and designing a good metric for a target task can be extremely difficult. The space in which the data is collected may not be suitable for the target task and chosen learner; it might be preferable to transform the space. Learning the distance metric to maximize performance on the target task has become an active area of research in machine learning. However, current state-of-the-art metric learning methods remain computationally very expensive and often cannot handle large datasets as we demonstrate in this paper.

Research on accelerating metric learning has yielded significant improvements. However, the majority of work is for method-specific improvements [3, 2]. While improving the efficiency has been a goal of recent metric learning work, our generic approach complements these recent method-specific improvements and for the first time makes large scale metric learning truly feasible.

Most metric learning methods require computing pairwise distances between training samples. This computational complexity grows quadratically in the number of samples making metric learning methods infeasible for large datasets. Our proposed method reduces the effective number of training samples by utilizing Locality Sensitive Hashing (LSH) based clustering. LSH [1] is a randomized algorithm for approximate nearest neighbor search. It is based on the idea that if two points are close then after partitioning by random projections these points will remain close. As we do not want large clustering overhead, LSH is an ideal candidate method for our purpose. However, LSH can map false positive samples into the buckets and reduces the performance of our method. We propose an approach for correcting LSH false positive using K-Means clustering inside the buckets with false positives.

Our approach achieves 1-2 orders of magnitude speedup, while maintaining comparable accuracy in all state-of-the-art metric learning methods used in our experiments. It outperforms baseline methods for training set reduction in accuracy and efficiency. Experimental results suggest that our method is useful for speeding up support vector machines and kernel learning methods while maintaining competitive accuracy.

Our paper makes three main contributions. First, we propose a generic LSH based clustering for fast metric learning by reducing the effective training set, thereby pushing the limits of feasibility for metric learning on large datasets. Second, we propose an approach for correcting false positives of LSH by using K-Means clustering inside LSH buckets. Third, we evaluate the robustness of our approach by comparing with the state-of-the-art alternatives. The results demonstrate our approach achieves more consistent accuracy across different training set sizes and superior accuracy when the compressed set is small.

2. REFERENCES

Measurement Theory in MOOCs via Topic Discovery

[Extended Abstract]

Jiazhen He
Department of Computing and Information Systems
The University of Melbourne
jiazhenh@student.unimelb.edu.au

James Bailey
Department of Computing and Information Systems
The University of Melbourne
baileyj@unimelb.edu.au

Benjamin Rubinstein
Department of Computing and Information Systems
The University of Melbourne
benjamin.rubinstein@unimelb.edu.au

Rui Zhang
Department of Computing and Information Systems
The University of Melbourne
rui.zhang@unimelb.edu.au

ABSTRACT
Massive Open Online Courses (MOOCs) have been widely studied recently in computer science, education and psychology. However, limited work have been explored from the cross-disciplinary perspective. This paper explores to use topic modeling, a machine learning technique, for measurement science and assessment in education.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data Mining

Keywords
MOOC, learning analytics, educational data mining, topic modeling, non-negative matrix factorization, measurement theory

1. INTRODUCTION
Massive Open Online Courses (MOOCs) have recently been the subject of a number of studies investigating the nature of learning in MOOCs, usually conducted from within a disciplinary perspective such as education, or psychology or computer science. Few studies have taken a cross-disciplinary approach. This paper explores to use a machine learning technique for measurement science and assessment in education. In particular, we explore to use topic modeling on forum discussion in MOOCs for measurement.

Measurement science in education and psychology is used to measure some intangible attribute, such as attitudes, abilities and intelligence. The key is to devise a set of items to measure the attribute of interest. According to Guttman scale, a basic and simple model for measurement, if the student responses on the items conform to Guttman scale, the items are accepted to have scaled unidimensionally to measure some attribute. A Guttman scale is present if a respondent who successfully complete any difficult item also completes all previous easier items; and once they get an item incorrect, they will also get incorrect on any item harder than that one. An example of Guttman scale measuring mathematical ability [1] is shown in Table 1, where the items are ordered in increasing difficulty, from Counting to Division.

Table 1: An example of Guttman scale measuring mathematical ability, where 1 means the person has mastered the item and 0 for not; the items are ordered in increasing difficulty, from Counting to Division.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P. 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P. 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>P. 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>P. 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>P. 5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Person 5 who has mastered the most difficult item Division, is expected to master all easier items (Counting, Addition, Subtraction and Multiplication) as well. The total score implies the person ability: the greater the higher. By knowing only the total score, the pattern of the item responses can be predicted, e.g. all the persons with total score 3 will have the same pattern 11100.

The items are often devised manually in education and psychology, which is human intensive. Furthermore, forum discussion is far more time-consuming to analyze but is nonetheless important. In this paper, we explore the automatic discovery of topics in MOOC forum discussion for measurement in MOOCs. The topics can be regarded as a set of items for measuring some latent skill if conforming to a Guttman scale. Since the response to an item in Guttman scale is dichotomous, in our case, the response of a student to a topic is whether contributing posts/comments on a topic or not. The goal is to generate a set of topics that make the student-topic matrix exhibit Guttman pattern (Table 1), which can be casted as an optimization problem. Based on Non-negative Matrix Factorization (NMF) [2], we automatically generate the items (topics) by designing a regularization mechanism to incorporate the education constraint (Guttman scale). Experimental results show that the topics generated by our algorithm yield a highly satisfactory scale for measurement.

2. REFERENCES
ABSTRACT
Anomaly detection in graphs has been the topic of interest in many application domains such as social and biological studies as well as road map networks and Internet security. The detection of anomalous events in such datasets can lead to a better understanding of vertex behavior. These data structures are often evolving over time, therefore, in this paper, we present a precise yet scalable approach for detecting anomalous graphs in dynamic graphs. We demonstrate an embedding scheme where the node vicinity is preserved while the time and space requirements are improved for analyzing large scale graphs.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—Data mining; G.2.2 [Discrete Mathematics]: Graph Theory

Keywords
Embedding schemes, Anomaly detection, Dynamic graphs.

1. INTRODUCTION
Anomaly detection corresponds to the task of finding interesting events that stand out in a collection of data points. The occurrence of such events leads to the observation of an anomalous collection of points with respect to the normal samples. These events can offer insights to the underlying data generation scheme. They often present actionable information and require further investigation. Such anomalies arise in many applications, namely health care, security, road map traffic, social networks and so forth. Anomalies in these applications may demonstrate significant incidents such as disease outbreaks, cyber attacks, road jams and fraudulent behavior.

The significance of this problem has led to a diverse collection of methodologies in this domain [3]. However, many of these techniques disregard the intrinsic dependencies between the data points; they propose techniques for handling independent or time-dependent data points while instances of data might contain inter-dependencies. Therefore, we need to devise approaches to deal with detecting anomalies in a much more complex settings. In addition to the inter-dependencies between the data points, we have to take time into account. The most widely known representations for encapsulating such rich data structures are graphs. However, due to the addition of a time component, we regard these data types as dynamic or time-evolving graphs [2].

2. DYNAMIC GRAPHS
Dynamic graphs are the most prevalent portrayal of vertex dependencies in real world networks such as social, collaboration, citation, road map and participatory sensing networks. We are interested in detecting emerging structures based on user interactions in such graphs. These structural variations can assist us in comprehending the dynamics of the user behavior.

The first step in analyzing such data structures is to develop a technique for managing the large scale of the input data. Many graphs such as social networks comprise of millions of users and billions of links. Therefore, our main focus is to reduce the graph dimensionality while preserving the node vicinity [4]. Therefore, we propose a dimensionality reduction scheme known as random projection for handling the large scale input graphs. Random projection is proven to preserve the Euclidean distance between data points within an $\epsilon$ of their original distances [1]. In regards to graphs, we can summarize node vicinity and structural properties in their spectrum. Therefore, we analyze the spectrum of the graphs after applying random projection and determine the influence of dimensionality reduction on the biggest eigenvalues of a graph and the spectrum in general. We applied random projection on Facebook wall posts and Enron email dataset. The results demonstrated that random projection can preserve inter-dependencies in a graph.

3. REFERENCES
Optimization and Programming Languages
Multiple Bug Spectral Fault Localization Using Genetic Programming

[Extended Abstract]

Neelofar
University of Melbourne
neelofar.eme@gmail.com

Lee Naish
University of Melbourne
lee@unimelb.edu.au

Kotagiri Ramamohanrao
University of Melbourne
kotagiri@unimelb.edu.au

ABSTRACT
Debugging is crucial for producing reliable software. One of the effective bug localization techniques is Spectral-Based Fault Localization (SBFL). It locates a buggy statement by applying an evaluation metric to program spectra and ranking program components on the basis of the score it computes. Recently, genetic programming has been proposed as a way to find good metrics. We have found that the huge search space for metrics can cause this approach to be slow and unreliable. Here we propose a restricted class of “hyperbolic” metrics, with a small number of numeric parameters. We show that effective metrics can be reliably discovered by finding parameter values of hyperbolic metrics using genetic programming.

Categories and Subject Descriptors
D.2.5 [Software Engineering]: Testing and Debugging—Debugging aids

Keywords
fault localization, program spectra, genetic programming

1. INTRODUCTION
Debugging software is a very important and resource intensive task in software engineering. Debugging requires fault localization at the initial stage. This is a tedious process, requiring substantial manual work. Due to this cost, many researchers are studying and proposing effective approaches which involve automated tools to aid fault localization. Spectral based fault localization (SBFL) techniques have gained much popularity in last few years due to their simplicity.

Performance of SBFL methods critically depend on the metrics used. Over one hundred metrics have been developed manually and evaluated for SBFL. More recently, genetic programming has been used to automatically develop metrics [2], resulting in many thousands being evaluated. For programs with a single bug, SBFL is relatively easy and we now have a good theoretical understanding [1]. The same is true for locating bugs which are “deterministic”. However, the general case where there are multiple bugs which may not be deterministic is much more complicated.

In this research we propose a new class of “hyperbolic” metrics which have a small number of numeric parameters whose values can be adjusted to vary the behaviour. Depending on the parameter values found by genetic programming, hyperbolic metrics can be optimal for single bugs, optimal for deterministic bugs or similar to other metrics known to perform well for some multiple nondeterministic bug benchmarks.

Rather than a single fixed formula, we use a formula with several additional numeric parameters. The general idea is to be able to pick different parameter values so the resulting formula performs well for different data sets. Using machine learning to find a small number of numeric parameters rather than a complete formula makes the learning task much simpler and, we have found, more efficient and reliable.

2. EXPERIMENTAL RESULTS
Table 1 shows comparison of average rank percentages of various metrics. The hyperbolic metrics out-perform other metrics in most of the runs and on average. Note that Kulczynski2 and Ochiai were the best performing metrics for two and three bug data. Thus beating these metrics is an impressive achievement. The research is submitted for publication and detailed methodology and results can be found at http://people.eng.unimelb.edu.au/lee/papers/hyp/

3. REFERENCES
Detecting Data Races in Java Programs

[Extended Abstract]

D A W Clarke
Dept. of Computing and Information Systems,
University of Melbourne
dclarke2@student.unimelb.edu.au

ABSTRACT
Multithreading is necessary to exploit the power of contemporary CPU chips. To avoid data races, accesses to variables must be guarded by critical sections. We describe a novel use of static analysis to identify critical sections, summarise this information across the classes in a system and, thence, detect data races.

Categories and Subject Descriptors
D2.4 Software Engineering - Software/Program Verification

Keywords
Java, data race, synchronization, volatile

1. INTRODUCTION
In the twenty years prior to the turn of the century, CPU clock rates increased from about 1MHz to about 1GHz. Contemporary CPU chips run at less than 4GHz. However, these chips are typically multi-core and have support for genuine simultaneous execution of multiple threads [1]. If an algorithm has poor performance, it is no longer possible to wait for the release of a faster processor. The only way forward is parallelism.

Multi-processing is an effective, but expensive way of achieving parallelism. Multi-threading is more efficient, but more prone to errors. Removing data races is an essential first step in diagnosing errors. We believe that our technique is an important aid to the development of multi-threaded algorithms.

2. FINDING DATA RACES
Data races occur when the same variable is read from and written to in different threads without synchronisation. They are avoided by confining access to shared variables within critical sections.

2.1 Identifying Critical Sections
Our algorithm begins by examining each method. We expand its control-flow structure into execution paths, reducing path explosion by using a "no-values" abstraction. We identify critical sections by searching the paths for the de-limiter patterns of known synchronisation techniques [2][3].

2.2 Summarisation
We derive a method summary as the set of critical sections that "must" occur, because they exist in all paths of a method, and those that "may" occur because they exist in only some of the paths. We form a system summary as the union of the method summaries for all the methods in a system.

3. EXPERIMENTATION
We have developed a subset prototype system in Java. We have tested it with a set of synthetic test programs to show that it correctly reports data races and identifies the relevant source line number (Figure 2 below).

3.1 Performance
We have conducted controlled scalability tests to show that the execution times of the detection tool will scale linearly with the size of the program being examined (Figure 3 below).

4. REFERENCES
The Benefits of Social Influence in Cultural Markets

Andrés Abeliuk
NICTA and UNIMELB
aabeliuk@student.unimelb.edu.au

Gerardo Berbeglia
NICTA and UNIMELB
g.berbeglia@mbs.edu

Manuel Cebrian
NICTA and UNIMELB
manuel.cebrian@nicta.com.au

Pascal Van Hentenryck
NICTA and ANU
pvh@nicta.com.au

ABSTRACT
Social influence has been shown to create significant unpredictability in cultural markets, providing one potential explanation why experts routinely fail at predicting commercial success of cultural products. To counteract the difficulty of making accurate predictions, “measure and react” strategies have been advocated, but finding a strategy that scales for very large markets has remained elusive so far.

Categories and Subject Descriptors
G.3 [PROBABILITY AND STATISTICS]: Stochastic processes

Keywords
Social Influence, Optimization, Dynamic Model

1. CONTRIBUTIONS
In this paper we propose a “measure and optimize” strategy to maximize expected profits in the market. Our main results are:

1. We show for the first time that dynamically showing consumers positive social information increases the expected performance of cultural markets.
2. We show our policy solves the resulting global optimization problem in strongly polynomial time, which means that the algorithm scales to large instances and avoid exploring the $n!$ possible rankings.
3. We show that, in reasonable settings, our policy does not introduce significant unpredictability and identifies “blockbusters”.

Overall, these results shed new light on the nature of social influence and how it can be leveraged for the benefits of the market. Our results contrast with earlier work which focused on showing the unpredictability created by social influence.

2. THE MUSICLAB MODEL
The descriptive model for the MusicLab [1] is based on the data collected during the actual experiments and is accurate enough to reproduce the conclusions in [2] through simulation. The model is defined in terms of a market with $n$ songs and is characterized by:

$\text{1. Its appeal } A_i \text{ which represents the inherent preference of listening to song } i \text{ based only on its name and its band.}$

$\text{2. Its quality } q_i \text{ which represents the conditional probability of downloading song } i \text{ given that it was sampled.}$

$\text{3. Its visibility } v_p \text{ which is the inherent probability of sampling a song in position } p.$

The MusicLab experiments present each participant with a ranking $\sigma$, where $\sigma_i$ denotes the position of song $i$ at time $k$ under two conditions detailed below. Under the social influence condition, the probabilities are given by $p_{i,k}(\sigma) = \frac{v_{\sigma_i}(A_i + D_{i,k})}{\sum_{j=1}^{n} v_{\sigma_j}(A_j + D_{j,k})}$, where $D_{i,k}$ is the number of downloads of song $i$ at time $k$. Under the independent condition, downloads are not shown to users, and thus the probabilities $p_{i,k}(\sigma)$ do not include the terms $D_{i,k}$.

3. RESULTS
We compare experimentally three different ranking policies. Download policy simply orders the songs by the number of downloads at each iteration. Performance policy maximizes the expected number of downloads at each iteration based on the probabilities $p_{i,k}(\sigma)$. Random policy presents a random ranking at each iteration. We annotate policies with either SI or IN to denote whether they are used under the social influence or the independent condition.

Figure 1: The Distribution of Downloads.
Figure 2: The Number of Downloads over Time.

4. REFERENCES

Download the program and proceedings from our website!


cis-dc@unimelb.edu.au
Tell Me What You Want And I Will Tell Others Where You Have Been

Anthony Quattrone, Elham Naghizade, Lars Kulik, Egemen Tanin
Department of Computing and Information Systems, University of Melbourne

Summary
We present an inference algorithm and demonstrate how it can effectively approximate original trajectories using solely the POI query results of a location-based provider returns from continuous user queries. Thus, highlighting potential privacy threats if query result data is exchanged amongst third parties.

Indirect Trajectory Inference Algorithm

<table>
<thead>
<tr>
<th>POI</th>
<th>r = 50m (%)</th>
<th>r = 100m (%)</th>
<th>r = 250m (%)</th>
<th>r = 500m (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>27.73</td>
<td>39.10</td>
<td>51.83</td>
<td>64.74</td>
</tr>
<tr>
<td>800</td>
<td>35.10</td>
<td>47.97</td>
<td>61.31</td>
<td>73.76</td>
</tr>
<tr>
<td>1,600</td>
<td>39.00</td>
<td>53.90</td>
<td>69.63</td>
<td>80.84</td>
</tr>
<tr>
<td>3,200</td>
<td>36.32</td>
<td>49.74</td>
<td>64.38</td>
<td>75.37</td>
</tr>
</tbody>
</table>

Experimental results using edge centrality

Conclusions and Future Work

Even coarse location information allows us to approximate a user’s trajectory within dense urban areas with high accuracy. Therefore, exchanging query results of LBS users instead of their tracks does not offer adequate privacy protection.

Privacy concerns (i.e., information directly shared by a location service provider) and adversary (i.e., an adversary could also have the information about all the POIs that were not revealed because they were not among the closest POIs)

Since the overall number of POIs is much larger than the number of POIs returned as a query result, the following Voronoi diagram may result in smaller cells, which may improve the accuracy of an inference attack algorithm. We are currently investigating these strategies.

References

Towards a Painless Index for Spatial Objects*

Rui Zhang*, Jianzhong Qi, Martin Stradling, Jin Huang

*Contact: rui.zhang@unimelb.edu.au

+ Patent Pending

**Aim**
A spatial index that achieves both easy implementation and excellent performance – Size Separation Index (**SSI**).

**Applications**
- Traditional spatial data applications.
- Modern location-based services, e.g., mobile apps that need to use a lightweight spatial database stored on a mobile phone.

**System Structure**

**Results**
- Standard: PostgreSQL’s B*-tree index
- R-tree: PostgreSQL’s R-tree index
- SSI: Proposed index

**Conclusions**
- SSI can be easily implemented on a DBMS.
  - SSI has been implemented on top of PostgreSQL and Oracle DBMSs.
  - SSI outperforms a DBMS integrated R-tree by up to three orders of magnitude.
- As a standalone implementation:
  - SSI outperforms other mapping based indices by orders of magnitude.
  - SSI is competitive with the R*-tree.
Analysis of Road Traffic Using Contrast Mining and GPS Trajectories
Xiaoting Wang, Christopher Leckie and Tharshan Vaithianathan

Introduction
Traffic congestion has been one of the most critical issues in many modern cities. A particular problem of interest is to characterize the traffic flow before and after an event, such as road closure, and assess the impact of this event. This study uses contrast mining techniques to identify emerging patterns like road closure in road networks. By computing the Growth Rate of traffic flows using vehicle GPS trajectories, road closure events can be detected and their impact can be characterized.

Contrast Mining
We calculate the traffic flow through the graph by aggregating the vehicle count on a sequence of road segments. We compute the Growth Rate of traffic of each sequence using:

\[
\text{Growth Rate} = \frac{\text{Traffic Flow After Event}}{\text{Traffic Flow Before Event}}
\]

Using the Growth Rate, we detect sub-networks of affected areas by computing the Local Outlier Factor (LOF) score, where

\[
\text{LOF}(x) = \frac{\text{Maximum reachability distance}}{\text{Average reachability distance}}
\]

A sub-network, or a sequence of road segments, will have a high LOF score when its Growth Rate deviates significantly from the usual levels of its neighboring networks. In Figure 2, the detected sub-networks are marked with blue triangles.

Results
We used a taxi GPS trajectory dataset in Beijing to identify and assess a road closure event during the 2008 Chinese New year (Figure 3). Edge #1 in Figure 3(b) was closed for 5 days. As a result, the traffic flow around #1 increased significantly (marked red in Figure 3(a)) while the flow into edge #1 reduced (marked green in Figure 3(b)).

We performed microscopic traffic simulation of a road closure event on George St in Sydney. Although the simulator still requires work, the resulting traffic flow in the presence of this disturbance shows similarity with the taxi trajectory data.

Conclusion
In this work we assessed the impact of road events using contrast mining and vehicle trajectories. Our next step is to build a predictive model of road traffic under various disturbances.
Online Medical Consultation (OMC): A review of literature, practice, and development of an evaluation model

Ibrahim Al-Mahdi1, Dr. Kathleen Gray2, Dr. Reeva Lederman1

1Department of Computing and Information Systems, 2Health and Biomedical Informatics Centre, The University of Melbourne

Background:
Online Medical Consultation (OMC) is the term used in this research to refer to internet-based patient-doctor (consumer-provider) medical consultations. OMC is an application of telemedicine which may found being more cost-effective delivery [15-17]. On the other hand, concerns have also been raised regarding quality and efficacy of these applications [6, 20-22]. A review of telemedicine papers stated that the quality and comprehensiveness of research in this area is still questioned [24]. Our exploration study in addition to previous literature indicates that further research is needed to investigate the various OMC dimensions [15, 25, 26].

OMC services can be seen as an intersection between the traditional telemedicine practices in terms of the clinical nature of the service, and internet services (e-services) in terms of the delivery over internet. Several reviews of e-services quality literature found that there is no agreement on measures for the e-services quality and there is no generic e-service quality model that would be applicable to any e-service [27, 28]. Therefore, we propose to evaluate OMC services utilizing an innovative model that combines e-service evaluation aspects together with the dimensions of clinical quality in order to capture the OMC’s unique features.

Academic reviews have cited numerous advantages of telemedicine consultation [12-14]. Academic reviews have cited numerous advantages of telemedicine consultation [12-14].

Methodology:
This PhD project aims at providing a guiding framework for evaluating OMC services to make it more efficient to undertake evaluation of OMC services on the same scale worldwide. The project forms a foundation work that uniquely defines OMC framework by combining concepts from e-services field and telemedicine, and by examining the phenomena from both providers and consumers perspectives.

OMC growth

OMC Countries

OMC Themes

Table 3. Consultation themes

<table>
<thead>
<tr>
<th>OMC themes</th>
<th>Percentage</th>
</tr>
</thead>
</table>
| OMC
| Consultations |
| Services |
| 75% of OMC services are cross-border |
| OMC services | 5 years since 2000 |
| OMC growth |
| Study 3: OMC evolution |

Methodology:
This research investigates OMC services from both providers and consumers perspectives. Multiple theoretical approaches will be utilized (technology adoption, e-service quality, and clinical safety), and mixed methods approach (qualitative, quantitative) will be followed in order to gain sufficient depth and rigor for the topic.

Importance of the study:
The significance of the proposed study is that it tracks at an early stage the emergence of OMC as a new approach to healthcare for providers and consumers. OMC is a rapidly emerging internet service phenomenon that started to attract giant providers like Google and Telstra [28, 29], which has so far gained little attention from researchers. OMC is a paradigm shift over traditional telemedicine and is likely to appeal to consumers who seek better choice or want the flexibility to see doctors out of regular office hours. This includes busy workers, parents of young children, travellers, and internet savvies in general.

References:
The Steiner Tree Problem

The problem
• Given a weighted graph, where nodes are either terminals or Steiner nodes find a sub-tree of the graph such that it spans all the terminals.
• Minimise the sum of the weights of the edges on the tree.

Why is it interesting?
• Which intermediate nodes to use? ⇒ NP-complete.
• Several applications in network design.
• Real-world applications have other “side” constraints (e.g.: force a minimum degree in some nodes, capacities on edges for some flow through the tree).
• Best algorithms use reduction to reduce the size of the problem which is not correct when there are side constraints.
• Only approximation algorithms exist for the variants of STP.

Contributions
• Use of Constraint Programming to solve the STP.
• Modelling two variants of STP and solving them to optimality.
• New tree propagator with explanations.
• New explanations for lower bounding techniques.

Search & Propagation

Search strategy
• The search engine will try to say whether an edge/node is in or out of the solution (Boolean decision variables) ⇒ the search builds a sub-graph that will be the solution tree.
• We start from the heaviest edges towards the lightest.
• In-domain-min: give the variables the value false first in the search, then try true.
• Propagation algorithms: infer values of variables when possible to help the search, detect failure.
• Learning: generate new clauses during propagation that used to remember inferences and explain failure so that mistakes are not repeated ⇒ accelerates the search.

Tree propagator
• Implemented with learning clauses.
• Break cycles.
• Find bridges and articulations: edges/nodes required to maintain connectivity.
• Find disconnected areas.

Steiner Tree propagator
• Based on the tree propagator.
• Additional lower bound.
• Learning on lower bound.

Lower bounding

Shortest paths: SPLB
• The sum of the paths between every pair of connected components is a lower bound.
• We compute the paths incrementally to extract the reason why the lower bound grows.

Linear program: LPLB
• Model the STP in LP (CPLEX12.4).
• Solve it to optimality and use it as a lower bound.

Variants of STP

Grade of Service Steiner Tree
• Each terminal node has a demand.
• Each edge has a capacity.
• Find a Steiner Tree of minimal cost such that each terminal A has at least one path reaching another terminal B where the minimum capacity in the path is larger than the demand at both terminals A and B.
• Used in telecommunication networks (e.g: video broadcasting).

Terminal Steiner Tree
• Find a Steiner Tree of minimal cost such that all terminals are leafs.
• Used in electronics (VLSI) where some electrical components must be leafs of the network.

Computational results

Comparison of our five different implementations in Chuffed (Splb, Lplb, Splb (n.l.), Sp+lp and Nolb) and an implementation in the Choco3 solver. We tested on three different sets of benchmarks (from 12 to 75 nodes).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SPLB</th>
<th>LPLB</th>
<th>Lplb (n.l.)</th>
<th>Sp+lp</th>
<th>Nolb</th>
<th>Choco3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nolb</td>
<td>12</td>
<td>14</td>
<td>101</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Sp+lp</td>
<td>12</td>
<td>22</td>
<td>112</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Lplb (n.l.)</td>
<td>15</td>
<td>13</td>
<td>99</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Lplb</td>
<td>329</td>
<td>534</td>
<td>3723</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lplb (n.l.)</td>
<td>352</td>
<td>1580</td>
<td>4636</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nolb</td>
<td>746</td>
<td>831</td>
<td>7016</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Choco3</td>
<td>4744</td>
<td>46175</td>
<td>67790</td>
<td>1.10</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>Splb</td>
<td>20879</td>
<td>63110</td>
<td>531033</td>
<td>22.25</td>
<td>22.25</td>
<td>22.25</td>
</tr>
<tr>
<td>Lplb</td>
<td>8097</td>
<td>9157</td>
<td>53731</td>
<td>5.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lplb (n.l.)</td>
<td>18000</td>
<td>37396</td>
<td>110711</td>
<td>6.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sp+lp</td>
<td>8882</td>
<td>10372</td>
<td>68844</td>
<td>7.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nolb</td>
<td>114538</td>
<td>192343</td>
<td>1460071</td>
<td>17.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choco3</td>
<td>8940257</td>
<td>8406288</td>
<td>152276771</td>
<td>3319.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Nolb = no lower bound was used
(n.l.) = no learning was used
Benchmarks taken from the SteinLib/OR-library.

Diego de Uña — Graeme Gange — Peter Schachte — Peter J. Stuckey
University of Melbourne
Department of Computing and Information Systems
TECHNOLOGIES AS AN INFORMATION SOURCE

EXPLORING FACTORS THAT IMPACT ON SAUDI FEMALE INTERNATIONAL STUDENTS’ USE OF SOCIAL TECHNOLOGIES AS AN INFORMATION SOURCE

Hara Dashti, Shahnaz Chang and Rachelle Bosua, The University of Melbourne

1. Research Question

What factors impact on Saudi female international students’ use of social technologies as an information source?

2. Research Gaps

Most Information seeking behaviour (ISB) studies focus ONLY on academic and library use needs (Chang et al. 2012, Alzougool et al. 2013). The importance of helping these students fulfil their information needs has been emphasized by many scholars (Alzougool et al. 2013, Sin & Kim 2013). Recent studies on international students' cross cultural adjustments identify ineffective information seeking or the lack of information sources as barriers to adapting to new cultures (Chang et al. 2012, Sin & Kim 2013). The importance of helping these students adapt to new cultures has been emphasized by many scholars (Alzougool et al. 2013, Sin & Kim 2013).

3. Research Question

What factors impact on Saudi female international students’ use of social technologies as an information source?

4. Proposed Research Model

A model derived from literature on ISB and ST use, this research in progress paper proposes the following propositions:

- **Proposition 1**: SFIS' information needs influence SFIS' use of online and offline information sources.
- **Proposition 2**: SFIS' information needs influence SFIS' use of online and offline information sources.
- **Proposition 3**: SFIS' information needs influence SFIS' use of online and offline information sources.
- **Proposition 4**: SFIS' information needs influence SFIS' use of online and offline information sources.
- **Proposition 5**: SFIS' information needs influence SFIS' use of online and offline information sources.

5. Proposed Data Collection Method

This working paper aims to propose a conceptual model of the factors that impact on SFIS' use of ST for their everyday life information seeking over time and across cultures. The model will be represented in DTPB. Therefore, in the coding process, concepts or themes represented in DTPB will be considered.

The use of ST for ELIS is also not entirely under the control of SFIS. Factors such as peer influence, attitude and behavioural control are expected to affect SFIS' intention to use Facebook and other social technologies. Given the fact that SFIS' ISB in general and via ST in particular are generally used in social contexts, Taylor’s & Todd’s (1995) Decomposed Theory of Planned Behaviour (DTPB) is selected as a theoretical framework to represent the relationships in the conceptual model. The condition satisfies the core assumption of the DTPB that the presence of constraints can inhibit both the intention to perform a behaviour and the behaviour itself.

6. Detailed Data Analysis approach

This paper proposes an approach to study the issue of ISB and ST use. From the above model we suggest the following propositions:

- **Proposition 6**: SFIS' ISB and ST use are influenced by their gender, age, academic level, and religious beliefs.
- **Proposition 7**: SFIS' ISB and ST use are influenced by their gender, age, academic level, and religious beliefs.
- **Proposition 8**: SFIS' ISB and ST use are influenced by their gender, age, academic level, and religious beliefs.
- **Proposition 9**: SFIS' ISB and ST use are influenced by their gender, age, academic level, and religious beliefs.

7. Conclusion

The results of this paper will be presented at the conference. The proposed model will be further tested and evaluated through empirical research. The results will provide insights into the factors that influence SFIS' ISB and ST use. This will help to improve the way SFIS use social technologies for networking, seeking, and sharing information.
Tour Recommendation on Location-based Social Networks

Kwan Hui Lim, Jeffrey Chan, Christopher Leckie and Shanika Karunasekera
The University of Melbourne and National ICT Australia
{limk2@student., jeffrey.chan@, caleckie@, karus@unimelb.edu.au

Motivation
- Unfamiliarity with Points of Interest (POI) to visit in foreign cities.
- Generic tours not appealing to the interests of users or adhering to their trip constraints.
- Time-consuming to plan an itinerary of multiple POIs (visit duration, visit order, etc).

Contributions
- Proposed the PersTour algorithm [1] for recommending personalized tours based on users’ interests and trip constraints.
- Proposed the TourRecInt algorithm [2] for recommending customized tours with a mandatory POI category based on user/tourist interests.
- Implemented a framework for extracting real-life user travel histories, which are then used for training and evaluation.
- Evaluated our algorithms on a Flickr dataset spanning multiple cities, against various greedy-based and random baselines.

Proposed Algorithms
- Based on a variant of the Orienteering problem [3], our main aim is to recommend a tour itinerary \( T = (P_1, ..., P_N) \) that:
  1. Maximizes POI popularity and tourist interest preference.
  2. Adheres to a distance/time budget.
  3. Starts and ends at specific POIs.
- Specifically, we proposed the PersTour algorithm [1] and TourRecInt algorithm [2].

Tour Recommendation Framework
- Our overall tour recommendation framework (see below) comprises the following steps:
  1. Map geo-tagged photos to a list of POIs if their coordinates differ by <100m.
  2. Construct the tourist travel history by connecting POI visits of the same tourist.
  3. Calculate POI popularity and tourist interests based on tourist travel histories.
  4. Use PersTour or TourRecInt algorithms to recommend customized tours.

Main Experimental Results
- PersTour and TourRecInt out-perform all baselines, based on tour popularity, interest, precision, recall and F1-score.
- For more information, refer to [1] and [2].

References

National ICT Australia (NICTA) is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

Research Excellence in ICT
Wealth Creation for Australia
CONCLUSION

This present study has both a descriptive and exploratory character since it aims to describe and organize information about the influence of BDA on information security process performance. The research addresses the following research question: How do big data analytics impact information security process performance?

We argue that compared to other organizations, companies with higher security process performance, have a substantial competitive advantage, a better reputation and market value and are more trustworthy. The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.

The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.

The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.

The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.

The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.

The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.

The information security field is a challenging one with many unsolved problems and numerous debates to solve these problems too. BDA has a productive area for the application of BDA techniques as it empowers the organization to monitor and optimally respond to security threats. It is easy to comprehend how BDA can be applied to information security, but in practice it becomes more challenging to envisage a direct link between BDA and ISM. The contribution of this research is that it integrates BDA with existing security intelligence solutions, organizations can develop security intelligence so that they can make accurate and timely decisions to respond to threats.

The role of BDA is to improve the performance of information security process. The use of BDA significantly improves the performance of information security processes. However, the understanding of how BDA can improve the performance of information security process remain unclear. This research argues that the main role of BDA is to increase information processing (IP) capabilities of an organization. However, such an increase of IP needs demands an increase in its IP capabilities. This would enable a more deliberate focus on the increase of IP capabilities in the processes where it is most likely to lead to the desired benefits.

The security data is treated as a moving target; as data grows, more sophisticated analysis methods (i.e., spreadsheets) from their databases; therefore, they have yet to harness the power of analytics particularly on their security data. Figure 2 illustrates the security analytics process.
Effective trust and reputation strategies
FRESCHI, Sergio, KIRLEY, Michael, and MILLER, Tim
Computing and Information Systems (CIS)
The University of Melbourne

Problem Statement

- Problem: How to efficiently identify who the most reliable or trustworthy individual is in highly dynamic online markets (e.g., stock markets) in order to preserve the (competitive) interests of their participants?

The motivation behind this project includes:

- To enrich participants’ future social interactions (e.g., cooperation, negotiation, etc.)
- To improve the individual member’s performance in terms of saving time and energy
- To enable better decision-making process (at run-time)
- Trust as one of the critical factors to enable information sharing in virtual organisations, thus increasing the efficiency and effectiveness of online services and communities

General View of Reputation Models

- Collect behavioural information about peers from different sources: direct interactions and/or external opinions or recommendations
- Scoring & Ranking (i.e., Trust metric)
- Decide which peer to interact with according to its trust and reputation values
- The actual transaction between the two participants is carried out
- Transaction outcome assessment (e.g., satisfactory/not satisfactory)

Open Challenges

- Open & dynamic systems: peers can join in or leave at any time and have dissimilar aims, no central authority, higher risk of unfulfilled commitments
- Heterogeneous participants or peers: for example, different behaviour (selfish, cooperative), abilities and preferences
- Dishonest Behaviour: peers telling lies about other peers (e.g., prejudice and bias, betrayal)
- Behaviour Change: from trustworthy to untrustworthy (some participants may change their behaviour throughout time)

Proposed Research Plan

At this early stage, the main goal of this research project is to investigate new possibilities of improving existing trust and reputation systems motivated by the current open challenges in this area as well as the critique to existing approaches found in the literature. Some possible ideas include:

- Combination of cognitive approaches with game theoretical ones.
- Time-sensitive and personalised trust metrics.
- Analysis of semantic domain information

3rd Annual Doctoral Colloquium // 22 July 2015
Introduction

The need for effective communication among team members is crucial for the success of any team. Communication challenges can arise due to differences in personal backgrounds, cultural disparities, or technological constraints. Understanding these factors and addressing them can significantly improve team performance.

Components of Effective Team Communication

1. **Face-to-Face Interaction**: Direct and immediate communication, allowing for instant feedback and clarification.
2. **Email and Instant Messaging**: Useful for asynchronous communication and documenting discussions.
3. **Video Conferences**: Enhances the personal touch of face-to-face interaction.
4. **Telephone Calls**: Simplistic and quick, suitable for urgent discussions.
5. **Social Media**: Ideal for informal communication and sharing updates.

Question

How do different components of communication affect team performance?

Components of the Socially Embedded Model Contribute Differently to Team Performance

Components of the socially embedded model have been identified to contribute differently to team performance. These components include:

- **Task**: The core work or project that the team is working on.
- **Reciprocal**: The mutual exchange of information and support among team members.
- **Team**: The collective unit of the team members.
- **Task**: The overall goal or objective of the team.

Results

Analysis of data from various studies has shown that the socially embedded model components significantly impact team performance. Effective communication strategies can be developed to enhance team performance, leading to improved outcomes.

References

[Insert references related to communication in teams and performance]

Conclusion

Communication is a critical aspect of team performance. Understanding and optimizing the components of the socially embedded model can lead to increased productivity and satisfaction among team members.
Better Health Explorer

Can’t find health information? Not sure what to read next? Try this out!

**Issues with current search engines:**
- People often have trouble thinking up the correct keywords for health issues
- Not specifically designed for health scenarios, resulting in a “trial-and-error” search process
- Unable to support the needs of the exploratory search approach.

**Main research question:**
How to design for health information seekers?

**Solutions with Better Health Explorer:**
- Recommendation for further readings on the current topic
- Summary and table of contents can help quick readers
- Coloured tiles show an overview of relevant information
- Play with sliders to discover more & observe the animated response
- No more keywords are needed!

We are evaluating this web app and we need your help!
Spend an hour with us, to help make searching for health information easier for everyone.
For information, visit [http://go.unimelb.edu.au/8oin](http://go.unimelb.edu.au/8oin) or scan this QR-code:

**Study Participants**
**!! REQUIRED !!**

Patrick Cheong-Iao Pang, Shanton Chang, Karin Verspoor, Jon Pearce
Department of Computing and Information Systems
Text Categorisation using Convolutional Neural Networks

Fei Liu
(Supervised by Timothy Baldwin and Trevor Cohn)

Dept of Computing and Information Systems
The University of Melbourne
fliu3@student.unimelb.edu.au

• CONVOLUTIONAL NEURAL NETWORKS
  - Successful and widely used in image classification
  - Achieves state-of-the-art performance on multiple NLP tasks

• RELATED WORK
  - Traditional approach: represent documents as bag-of-words vectors;
    problems: data sparsity, ignore word order
  - Convolutional neural networks:
    - Make use of dense, continuous word representations
    - Capture internal structure of data (e.g., word order)
    - Handle documents of arbitrary length (dynamic k-max pooling)

• PROPOSED APPROACH
  - Represent words by dense, continuous vectors from word embeddings
  - Obtain document vector representations using convolutional neural net:
    - Wide convolution: dot product of a vector of weights in \( \mathbb{R}^m \) and a
      vector of inputs in \( \mathbb{R}^s \)
    - Dynamic k-max pooling: select the \( k \) most active features in the input
      at layer \( l \), \( k_l \) is determined dynamically

• DOCUMENT CATEGORIZATION
  - Classify documents based on their vector representations
  - Represent words by dense, continuous vectors from word embeddings

• EXAMPLE DATASET
  - Reuters Corpus Volume 1:
    | Label | Train | Test | Classes |
    |-------|-------|------|---------|
    | Single | 15,564 | 49,838 | 55      |
    | Multiple | 23,149 | 78,126 | 103     |

Figure from Kalchbrenner et al.
Towards a Model of Information Security Management Practices (ISMPs) in Organisations

Abstract
There is growing recognition of the role that management performs in protecting organisational information. However, our review of the academic and professional literatures did not find an empirically sound and coherent view of the range of management activities that can be applied as part of an information security program. As a result, organisations have insufficient guidance on what methods can be implemented to meet security objectives. Further, organisations have no empirically evidenced benchmark against which management practices can be assessed. This research project aims to develop a rigorous, comprehensive and empirically evidenced model of information security management practices (ISMPs) to provide organisations with comprehensive guidance.

Background
The threat of leakage of trade secrets and intellectual property, disruption of mission-critical systems, and malicious attack from both insiders and outsiders makes information security a high priority for organisations. Although technical security controls have always played a critical role in reducing security risk exposure, recent research has highlighted the critical role of managerial controls in the pursuit of security objectives (e.g., see Knapp & Ferrante 2012; Ahmad et al. 2012; Lim et al. 2012).

Prior research in information security management (ISM) focuses on a variety of issues such as: the importance of security policy, development of good security policy, use of deterrence theory to investigate employees’ compliance with security policy, risk management, incident response and security awareness. However, it is argued that much of this research is disjointed and does not provide comprehensive guidance to organisations about information security management practice (Qingxiong et al. 2008, Lim et al. 2012).

The lack of a comprehensive model that provides sufficient guidance on the range of managerial practices of information security has serious consequences for the modern organisation given the significant exposure to security threats. Organisations embarking on an ISM program have insufficient guidance on what methods can be implemented to meet security objectives. Further, organisations have no widely accepted benchmark against which improvement in existing management practices can be assessed.

Aim
This research project aims to develop a model of ISMPs to address the need to identify, classify and understand the management practices of information security. The model will provide organisations with a means to understand, implement and assess ISMPs in a systematic and comprehensive manner in order to enable them to protect themselves from a wide range of threats.

Significance
The proposed model provides comprehensive guidance on ISMPs that can be implemented and what activities can be performed to deliver effective security management. Additionally, practitioners will be able to benchmark their information security management activities against the model.

Research Design
This research is exploratory and follows a qualitative research design. A preliminary model of ISMPs has been developed from the literature. The model is to be refined and validated by information security experts. First, a set of 15-20 semi-structured interviews with information security experts will be conducted with the aim of inviting comment on the preliminary ISM model for the purpose of refinement. Secondly, a set of 4-6 focus groups consisting of 6 security experts will be used to perform the final validation and refinement of the model.

Research Model

References
Knowledge Base Population over Biomedical Data

P.C. Nagesh, Kotagiri Rao, Trevor Cohn
Computing and Information Systems, University of Melbourne

Introduction

- Biomedical data is huge and growing fast
- Scientific literature such as MEDLINE has about 18M articles and growing at around 700K articles every year
- Non textual data such as MRIs growing at around 10 PB every year
- Experimentation or retrials are costly and (re)discovering information from prior research is crucial
- Mining of this data has gained tremendous importance owing to critical applications
  - Accelerated Drug Discovery
  - Protein function prediction
  - Predicting trends related to societal health or outbreaks
- Our goals:
  - Knowledge base population of Biomedical data
  - Real time knowledge detection while authoring the Biomedical literature

Entity-Relations in the BioMedical domain

Key Computational Techniques

- Unsupervised Entity-Relation extraction from linguistic analysis [1]
  - Infer entities from noun phrases and relations from verb phrases with other linguistic cues.
  - Integrate matches with external dictionary when available
- Distant Supervision for creating training data
- Optimal selection of training examples for manual annotation [3]
- Bootstrapping with few annotated samples
- Cost augmented SVMs to deal with unbalanced training data

Challenges in BioMedical Text Analysis

- Biomedical text is significantly unique and general NLP tools such as POS taggers dont work well
- Indirect performance evaluation that is not only dependent on the few annotated samples
- Integration across diverse knowledge sources
- Better ML models or learning algorithms
  - Distant supervision produces unbalanced and noisy training data
  - Bootstrapped learning leads to semantic drift that is hard to detect or avoid
- Intelligent Information Integration
  - Provenance of facts
  - Infering insights from multiple facts
  - For example, effective inference in medical domain requires factoring in prior history of patient in addition to current symptoms and medication.

Resources [2]

- MetaMap: Open source software for identifying UMLS concepts in text
- Biological NERs (ABNER, BANNER) built on Conditional Random Fields
- ORBIT: Online catalog of software tools and electronic documents in Biomedical domain
- Stanford Deep Dive: Entity extraction software and Biomedical datasets

BioMedical literature mining

Research Directions

- Domain specialized syntax and semantic tools
  - Medical literature is characterized by heavy use of conjunctions that challenge phrase chunking
  - Synonyms and abbreviations are context and topic specific. For example “R.A” can refer “Rheumatoid Arthritis” in some context and “Renal Artery” in other.
- Whole text analysis
- Open Entity Relation extraction
- Deep Question-Answer systems
- Graphical Models for Biomedical text mining
  - Probabilistic Graphical Models such as CRFs need large amounts of training data
  - Calibrating PGMs for unbalanced training data is hard

Applications

Ref: The PharmGKB Knowledge Pyramid.

References

Unsupervised entity-relation analysis in IBM Watson.
In Proceedings of the Third Annual Conference on Advances in Cognitive Systems ACS,
page 12, 2015.

Biomedical text mining: a survey of recent progress.

Medical relation extraction with manifold models.
Introduction

Biomedical pathways e.g., gene regulatory, and signal transduction networks are widely used to describe processes of gene-protein or protein-protein interactions, which could explain specific health conditions in the biomedical and pharmaceutical research. Since the scientific literature contains knowledge about relationships and events involving biomolecules such as proteins, genes, and chemicals, approaches for extraction of biomolecular entities and events have been developed. However, coreference expressions such as pronouns (e.g., it, they), and definite noun phrases (e.g., the protein, these genes) are one of major obstacles for existing methods, limiting system performance in information extraction from biomedical literature to sentences that directly mention entities [1]. Abundant anaphoric mentions are used to refer entities that were previously mentioned in the same text, such as when interactions or events are written over successive sentences in the biomedical context. Those relationships are ignored by most event extraction systems.

The BioCreative V (2015) shared task Track 4 addressed the task of extraction of causal network information in terms of the Biological Expression Language (BEL), a formalised representation language for biological expression. It aims to simulate development of tools that extract events, and generate a corresponding BEL statement. To address the task, we developed a pipeline system which consists of an event extraction system (TEES) [2], an automatic system for generating BEL statements [3], and a coreference resolution module. We investigate how incorporating coreference resolution might impact event extraction in the biomedical domain.

Background

Biological Expression Language (BEL) statement

- Knowledge representation of relationships between two BEL entities
- Triple structure of subject - predicate - object
- Two classes of biological entities (Terms)
- Use of external terminologies and ontologies (Namespace)

Methods


Input text: Regulation of the stability of p53 is key to its tumor-suppressing activities. mdm2 directly binds to the amino-terminal region of p53 and targets it for degradation through the ubiquitin-proteasome pathway. (PubMed 11278372)

CR Identification: Regulation of the stability of p53 is key to its tumor-suppressing activities, mdm2 directly binds to the amino-terminal region of p53 and targets it for degradation through the ubiquitin-proteasome pathway.

CR Substitution: Regulation of the stability of p53 is key to its tumor-suppressing activities, mdm2 directly binds to the amino-terminal region of p53 and targets it for degradation through the ubiquitin-proteasome pathway.

TEES: Positive Regulation T8 (targets) Cause T2 (mdm2) Theme E3 (Protein_catabolism: p53 degradation)

BEL Generator: p(MGI:Mdm2) -> deg(p(MGI:Trp53))

Figure 1. Pipeline system Architecture with example

Results

Table 1. Statistics of coreference relations in the Gold standard corpora

<table>
<thead>
<tr>
<th>Coreference Types</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset</td>
</tr>
<tr>
<td>Relative pronoun</td>
<td>1,313</td>
</tr>
<tr>
<td>Personal pronoun</td>
<td>257</td>
</tr>
<tr>
<td>Possessive pronoun</td>
<td>506</td>
</tr>
<tr>
<td>Definite noun phrase</td>
<td>74</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2,186</td>
</tr>
</tbody>
</table>

Described in Table 1, we analysed the gold standard datasets regarding to types of coreference, and only personal pronouns are substituted with their antecedents in the component of coreference resolution substitution.

The coreference resolution process, our system produced 74 BEL statements in 244 sentences on the training dataset, while 66 BEL statements were generated without the coreference resolution module (Table 2).

Table 2. Statistics of BEL statement output by approaches of non-coreference, and coreference resolution on the gold corpora

<table>
<thead>
<tr>
<th>Numbers of BEL Statement</th>
<th>Non-Coref. Resolution</th>
<th>Coref. Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training dataset</td>
<td>66</td>
<td>74</td>
</tr>
<tr>
<td>Test dataset</td>
<td>58</td>
<td>64</td>
</tr>
</tbody>
</table>

We evaluated our system including the coreference resolution module at Term, Function and Relation level on the Training dataset comparing with non-coreference resolution module in Table 3.

Table 3. Comparison of overall performance between non-coreference resolution, and coreference resolution approaches on the Training dataset

<table>
<thead>
<tr>
<th></th>
<th>Non-Coref. Resolution</th>
<th>Coref. Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>51.9 / 55.4 / 53.4</td>
<td>59.2 / 56.3 / 57.7</td>
</tr>
<tr>
<td>Function</td>
<td>33.3 / 5.4 / 9.3</td>
<td>37.5 / 12.0 / 18.2</td>
</tr>
<tr>
<td>Function-Secondary</td>
<td>33.3 / 17.4 / 21.8</td>
<td>42.9 / 15.8 / 23.1</td>
</tr>
<tr>
<td>Relation</td>
<td>16.0 / 18.2 / 17.0</td>
<td>27.4 / 27.0 / 27.2</td>
</tr>
<tr>
<td>Relation-Secondary</td>
<td>58.3 / 63.6 / 60.9</td>
<td>64.7 / 69.8 / 67.2</td>
</tr>
</tbody>
</table>

Conclusions

In this study, we have developed a pipeline system for biological event extraction, and BEL statement generation that incorporates a coreference resolution component. We have demonstrated that coreference resolution is helpful to improve performance for event extraction in biomedical literature. For future work, we will further expand the scope of the coreference resolution system to resolve relative pronouns and definite noun phrases, in order to achieve further improvement in our information extraction capability.

References

The Voluntary Use of Devices

Individuals are undertaking extreme body modifications with electronic devices. There are several examples of individuals who voluntarily insert microchips or other electronic devices under their skin. However, this arena has had little academia research; the majority of the available literature is not peer reviewed [1] or appears in opinion pieces [2] and news articles [3-5]. This project explores individuals with voluntarily inserted devices, excluding devices specifically designed for medical purposes (e.g., pacemakers).

So far ten interviews have been conducted. Preliminary results show insertables fall into three categories:

1. RFID chips
2. NFC chips
3. Neodym magnets

The possible locations of these devices are seen in the diagrams:

This research introduces a new class of devices: Insertables - devices that are inserted through skin or implanted under skin. Insertables have evolved from wearables, which in turn evolved from luggable devices.

Interviews are still being conducted and the study will explore:

A - What are the characteristics of these individuals?
B - Why did they decide to modify their bodies instead of using wearables?
C - Which capabilities of the device do they wish to exploit?

Acknowledgements
I thank my supervisors Associate Professor Frank Vetere and Associate Professor Shanton Chang. I would also like to thank the participants of study one so far for their time.

References
[3] Cheer, L. (2015) Australian man who’s had a microchip inserted into his hand so that he can do more with the iPhone 6...maybe, Daily Mail.
Using Community Question-Answering Data to Answer User’s Questions

Doris Hoogeveen
Supervised by Tim Baldwin and Karin Verspoor
Dept. of CIS, University of Melbourne, Australia
dhoogeveen@student.unimelb.edu.au

Introduction

Task: Identify questions in community question-answering archives that are semantically similar to new questions.

Motivation: The complex questions that people post on CQA websites are difficult to answer using search engines.

Assumption: The answer to an archived question that is identified as semantically similar to a new question will satisfy the information need of the user posting this new question.

Example CQA thread

Challenges

• The lexical gap problem: semantically similar questions often do not have any lexical overlap.
• The size of the data set: +/- 1.3 million questions. This makes pairwise comparison infeasible.
• Evaluation: what to do with new questions for which there are no similar questions in the archived set?

We need a way to quickly filter out a large portion of the non-similar question pairs so that the data set becomes manageable and we can use more complex methods without losing efficiency.

Methods

The following methods have been investigated for the filtering task:

• Simple heuristics (TN: 12.8%, FN: 4.2%)
• Retrieval models (TN: 95.3%, FN: 18.2%)
• Near-duplicate detection (TN: 99.1%, FN: 32.6%)
• String similarity measures (too slow)

Things to investigate in the future:

• User modeling
• Classification using both content and structural features
• Approximate matching in IR
• Deep learning: doc embeddings, RNNs

Related Work

• Jeon et al. (2005): applied a monolingual word-based translation model to retrieve archived questions that are a “translation” of the input question.
• Xue et al. (2008): translation-based language model (TBLM). They combined Jeon et al. (2005)’s model with Ponce et al. (1998)’s language model to retrieve duplicate questions.
• Madnani et al. (2012): using machine translation evaluation scores as feature values in a classifier to identify paraphrases.
• Socher et al. (2011): Recursive auto-encoder for paraphrase identification.

Annotated Data

• Users of the forum can flag questions as begin duplicates of earlier posts.
• Users can upvote or downvote answers.
• Users can get points for providing good answers.
Query Processing on Time-evolving Graphs

Masoomeh Zameni¹, Christopher Leckie², Masud Moshtaghi³

Department of Computing and Information Systems
University of Melbourne

**INTRODUCTION**

**Problem:**
- Real graphs have millions of nodes and billions of edges
  - Urban traffic data
  - Social networks
  - Web graphs
- These graphs frequently change over time (Time-evolving graphs)
  - Appearing or disappearing nodes
  - Appearing or disappearing edges
  - Fluctuations in edges weights

**Objective:**
- Compressing time-evolving graphs while:
  - Maintaining most of their intrinsic structural patterns
  - Effectively preserving the results of graph queries (e.g., shortest paths (SP) queries)

**Importance:**
- Enabling the analysis of large graphs
- Saving the cost of storage
- Reducing the cost and time needed for processing

**DEFINITIONS**

**Time-evolving graph:**
- A triple $DG = (V, E, g)$
  - $V$: a finite set of vertices or nodes
  - $E$: a set of edges, an unordered pair $\{u, w\}$ of distinct nodes
  - $g$: a function which maps weights to edges

**Example:**
- Traffic network
  - Nodes: Intersections
  - Edges: Road segments connecting intersections
  - Edges weights: The number of cars traveling on a road segment

**CHALLENGES**

- Analysing the large volumes of data
- Preserving the dynamics of the graphs
- Retaining the time information in the graphs
- Balancing query accuracy and compression ratio

**METHODOLOGY**

**Idea:**
- Leaving just enough information to answer the queries
- Applying the idea of video compression algorithms based on BPI frames
  - Key frames (I)
  - Forward-predicted frames (P)
  - Bi-directionally predicted frames (B)

- I-frames: are the least compressible but don't require other video frames to decode.
- P-frames: can use data from previous frames to decompress and are more compressible than I-frames.
- B-frames: can use both previous and forward frames for data reference to get the highest amount of data compression.

- We can consider each graph as an I, B, or P frame

**Solution:**

**Phase 1:** Partitioning and extracting I-frames
- Grouping similar snapshot graphs together
- Extracting two representative graphs from each cluster
  - $G_c$: The intersection of all snapshots in a cluster
  - $G_u$: The union of all snapshots in a cluster

**Phase 2:** Shortest Path Query Processing based on I-frames
- For any nodes $u, v$ and snapshot graphs $G_i$ in cluster $C$: $SP_i(u, v) = SP(u, v) \leq SP(\cap G_i(u, v) \leq SP(\cup G_i(u, v))$

**CONCLUSION**

- Representing real networks by large graphs
- Capturing the dynamic changes of these graphs by a sequence of time-evolving graphs
- Describing a method which efficiently answer SP queries

**REFERENCES**


**CONTACT**

¹ m.zameni@student.unimelb.edu.au
² caleckie@unimelb.edu.au
³ masud.moshtaghi@unimelb.edu.au
Implementing the greedy Set Cover algorithm efficiently for disk-resident data

LIM Ching Lih (MPhil candidate), Alastair Moffat and Tony Wirth (Supervisors)

The Set Cover problem

An instance of SET COVER consists of a universe of elements, \( U = \{e_1, e_2, \ldots, e_n\} \) and a family of sets, \( F = \{S_1, S_2, \ldots, S_m\} \), with \( S_i \subseteq U \) and \( \bigcup_{i \in F} S_i = U \). The problem is to find the smallest subset \( A \subseteq F \) such that the union of all sets in \( A \) covers \( U \) as a whole [1, 2].

Figure 1: An instance of the problem, where \( U = \{e_1, e_2, \ldots, e_7\} \) and \( F = \{S_1, S_2, \ldots, S_7\} \).

Computing the smallest (optimal) solution to an instance would take at least \( O(2^n) \) time, making it impractical for datasets of very small size.

The benefits are:
1. Reduction of the search area on remaining data for exact and approximation algorithms.
2. Finding of sets that are part of the optimum sooner.
3. The precomputation time is in \( O(M) \).

Background of our approximation algorithm

One well-known polynomial-time algorithm can produce an approximation solution that is within a factor 1 + \( \ln(n) \) of the optimum in size [3, 4]. It works by greedily and repeatedly adding sets with the largest numbers of uncovered elements to \( A \) until \( U \) is covered. It may be implemented as follows:

1. Use a priority queue to assign a priority to each set, based on its uncovered elements.
2. Pick and add the highest priority set to \( A \) (repeatedly until \( U \) is covered)
3. Update the sizes of the remaining sets

For an input of size \( M \), there are maximum \( M \) updates on sets, and each update accesses to a particular set at any time. This random access is relatively expensive on a slow disk.

Our disk-efficient variant, SEMI-EXTERNAL LAZY GREEDY (SELG), works as follows:

1. Divide an input into a hierarchy of buckets of set-size ranges
2. Process one bucket by one bucket in order of decreasing set sizes
3. Retract one set from the bucket (repeatedly until \( U \) is covered)
4. Add the set to \( A \) if it does not contain any covered element
5. Otherwise, cleanse and reinsert the smallest set into one appropriate bucket

Choosing the bucket quickly

The binary representation of a number can be exploited to simplify the whole management of buckets. A single word of space and one array of length \( \log n \) would be sufficient.

Figure 2: Binary representation of the bucket address.

Figure 2 shows how a maximum priority of 20 represents the bucket. The two 1-bits at binary positions 4 and 2 represent buckets of width 16 and 4, with priority ranges \([16, 20]\) and \([12, 16]\) respectively. An array \( B \) "expanses" the binary representation by corresponding each binary position to the cell position. The patterned cells are unused because there are no corresponding buckets in the maximum priority. Wherever the priority is decreased, each corresponding bit implicitly determines the existence and size range of a bucket.

Moreover, the location of a particular bucket is computed in \( O(1) \) without accessing the bucket list. The corresponding size of the bucket undergoes one bitwise operation and is compared against the maximum priority in bits, using another bitwise operation.

Hapax legomena

A hapax legomenon is a word that occurs only once in a single text.

Lemma 1. If a SET COVER instance, sets contained hapax legomena are in the optimal solution.

Proof. If every element in the universe is to be covered, all sets with hapax legomena must be in valid solutions. Otherwise, these solutions would not cover every element in \( U \).

The following process, HXL, of identifying such sets and outputting a reduced instance of remaining sets and elements is described:

1. Create a counter array for \( n \) elements in \( U \), with the counters being zero.
2. For each set \( S_i \) in an input, increment the corresponding counter of every element in \( S_i \) except the counter of each element \( e_i \) in \( U \).
3.1. If the counter is 1, then \( e_i \) is a hapax legomenon
3.2. Add the set \( S_i \) containing \( e_i \) to the growing solution \( A \) and cover its elements
4. Create a new instance \( F' \) of remaining sets and uncovered elements
4.1. Remove the remaining sets and elements in a new sequence of continuous integers
5. Call an external algorithm on \( F' \) to compute a solution \( A' \)
6. For each set in \( A' \), add the corresponding set of the original input to \( A \).

The last pass counts the occurrences of the elements in the input, the second pass reads eligible sets that contain hapax legomena, and the third pass consolidates the remaining sets and elements into a new instance. Since each pass examines at most \( M \) elements of the input, HXL executes in \( O(M) \) time.

The virtues are:
1. Reduction of the search area on remaining data for exact and approximation algorithms.
2. Finding of sets that are part of the optimum sooner.
3. The precomputation time is in \( O(M) \).

Better solution and shorter running time

In experiments, the instances described in Table 1 are solved without HXL. Then, the same inputs are precomputed first and the resulting reduced instances are solved then.

Table 1. Properties of the datasets in our experiments. The term "HL" corresponds to a hapax legomenon.

Table 2 and 3 measure our algorithms. The disk-oriented algorithm, DFG, serves as a baseline because it is fast and has a similar system of grouping sets by size ranges [5].

Table 1. Properties of the datasets in our experiments. The term "HL" corresponds to a hapax legomenon.

Table 2: Execution times (seconds) of the algorithms on the instances, and size percentages of the instance reductions.

Table 3: Sizes of the solutions computed in the same experiment runs.

The combination of HXL and SELG takes a shorter time than SELG alone on the last four instances. Its solutions are smaller than the ones of either SELG or DFG. The input-size reductions are exceeding 79% of their initial sizes. Only the MUSHROOM instance is unsuccessfully searched for hapax legomena.

Conclusion

The binary representation of a maximum set size allows a compact management of buckets that changes automatically whenever this size is decreased. The bucket lookup uses two bitwise operations and one word of space to find a specific bucket in constant time.

The HXL precomputation identifies sets, which have one-occurrence elements, and adds them to a partial solution in \( O(cM) \) for constant \( c \). It then removes these sets and their covered elements from the original input, and passes remaining sets and elements to an algorithm to solve. This decreases both running time and size of the final solution to an instance.
Is greed good?
Towards sustainable and profitable resource harvesting
Friedrich Burkhard von der Osten
Computing and Information Systems (CIS)
The University of Melbourne

1. The Problem

- The Common Pool Resource Game (CPRG) [1] models resource harvesting by a group of individuals.
- It is a Social Dilemma, where conflicting goals have to be balanced.
  1. The individual goal is to make a profit.
  2. The global goal is to maintain the resource.
- The dynamics of the game change with environmental conditions; there is no fixed solution.

2. Cooperation Models In The CPRG

As self-governance is desired to solve the CPRG, literature suggests various mechanisms:
- Behavioral modifications, such as norms, conformity, trend and expectations of behavior of other individuals.
- Changing the context of the game by means of participation (less beneficial for the poor) and the perception of external regulation or group composition.
- A promising approach is punishment, but it is costly.
- Common factor is the necessity for information, such as communication (exchange), experience (accumulation and refinement) and information from multiple sources (accuracy and availability).

Problems:
1. Many existing solutions focus on sustainability, not profitability.
2. Not enough focus on information playing a crucial role.

3. Why Modified Learning Can Help

- Humans use learning to solve tasks with dynamic solutions [2].
- No ground truth required, use environmental response.
- No model of environment needed.
- Online adaptation (balancing exploration vs exploitation)
- Bottom-up perspective models self-governance.
- Multi-Agent Reinforcement Learning can make use of Information Sharing.

4. Model

- CPRG model is a set of differential equations defining resource and harvesting behavior:
  \[
  N_t = N_{t-1} + G(N_{t-1}) - B(N_{t-1}, N_t) \tag{1}
  \]
  with resource \(N\), growth \(G\) and harvest \(B\) depending on cumulated invested effort \(X\).
- Q-Learning is used to develop individual action-selection policies using the environmental response; action-value pairs are evaluated with a corresponding reward:
  2. Profit (individual response).
- Information sharing across multiple games to support Q-Learning, as individual action response differs with large groups.
  1. Information from multiple different sources is used.
  2. Separate action-selection policies are developed and shared together.

5. Sample Results

Sample results show that multiple pools sharing information can sustain larger groups of agents (left) and agents adapt well to environmental conditions such as a change in resource growth (right).

6. Conclusions

Extensive simulations with the model have yielded the following insights:
- Reinforcement Learning and Information Sharing allow agents to harvest sustainably and profitably.
- Information Sharing improves stability and learning speed.
- More information shared in parallel leads to higher profits and resource levels.
- Agents can adapt well to environmental changes.
- With gradual changes agents can adapt to more extreme conditions.
- Too large profit margins facilitate greed and overexploitation, too low profit margins lead to economic losses.
- Larger groups have a detrimental effect on sustainability.

7. Open Challenges

The current results prompt a range of directions for further research:
- A new learning paradigm involving predictions and analysis of consecutive actions is explored to further improve agents’ adaptability to changing environments and reduce the initial cost and duration of learning.
- The model can be extended by the Theory of Mind aspect, meaning agents will have a model to estimate other agents’ behavior.
- Cross disciplinary projects will help establish a translation of the results into real-world settings and develop psychological models to make them applicable.

8. References

We examine a context where the indigenous members living in the diaspora (Australia), wish to connect with members in situ (Kenya), in order to maintain their indigenous identity while in their place of migration (Australia). Thus, we seek to understand how technology can mediate the nurturing of indigenous knowledge between in situ and diaspora indigenous community members.

Taking seriously the design themes below, our next step is to investigate how social, situated and tangible technologies can mediate IK learning environments. We will do this by building a prototype that addresses the challenges raised in these proposed design themes. Our prototype will enable us to evaluate how technology can mediate learning environments for participants who are remotely located, and who apply an indigenous lens to knowledge.

**Supporting bodily literacy**
Indigenous knowledge is not only expressed verbally, but also through non-verbal sounds, ululations, silences, changes in tone and resonance, mime, and body and gestural movements that constitute indigenous languages.

**Not translating**
While a lot of work in HCI focuses on translation for multilingual contexts, we found opportunities not to translate. Speaking a different language allowed the teachers to correct each other privately, while amidst the students.

**Illuminating knowledge in artefacts**
Artefacts hold stories and memories. The role of artefacts is key in indigenous environments, giving an opportunity for technology to animate the stories and memories that indigenous artefacts hold.

**Enabling apprenticeship through space**
Apprenticeship is one of the most common teaching methods in indigenous contexts. The challenge here is for technology to enable apprenticeship to occur when the teachers are located on the ancestral land (in situ), and the students live in the diaspora.
Understanding Knowledge Leakage & BYOD: From a mobile worker’s Perspective
Carlos Andres Agudelo Serna
cagudelo@student.unimelb.edu.au
Supervised by: Dr. Rachelle Bosua, Dr. Atif Ahmad and Dr. Sean Maynard

References

Phase 1 of the case study:
The groups analysed this data using their own in-house approaches and found extremely diverse (incomparable!) results (i.e., variants).

Phase 2 of the case study:
The same groups were asked to use a common bioinformatics environment: the Galaxy platform made available through the Genomics Virtual Laboratory (www.genome.edu.au) running on the NeCTAR Research Cloud (www.nectar.org.au). This phase focused on attempting to support reproducibility and comparison of the bioinformatics analysis systems with a simpler data set (6 exomes from individuals affected by DSD). The teams used a range of Cloud-based tools in undertaking this activity. Encryption technology was used to ensure the privacy of the data on the Cloud.

In this context, the NeCTAR-funded Endocrine Genomics Virtual Laboratory (www.endovl.org.au) has explored targeted workflows on the Cloud. We illustrate this through case studies exploring a rare genetic disorder - DSD and the different approaches that scientific teams adopt. We identify that improved consistency of results is urgently needed for translation into a clinical setting.

INTRODUCTION

METHODS

A two-phase case study was conducted as part of the endoVL, focused on association of phenotypic databases with the variants that might be involved with cause particular kinds of DSD. The NGS data for a large family of 15 from Indonesia, 7 of whom had DSD, was made available from clinical collaborators from the Royal Children’s Hospital in Melbourne and subsequently sent to three separate bioinformatics groups across Australia for analysis.

Phase 1 of the case study:
The groups analysed this data using their own in-house approaches and found extremely diverse (incomparable!) results (i.e., variants).

Phase 2 of the case study:
The same groups were asked to use a common bioinformatics environment: the Galaxy platform made available through the Genomics Virtual Laboratory (www.genome.edu.au) running on the NeCTAR Research Cloud (www.nectar.org.au). This phase focused on attempting to support reproducibility and comparison of the bioinformatics analysis systems with a simpler data set (6 exomes from individuals affected by DSD). The teams used a range of Cloud-based tools in undertaking this activity. Encryption technology was used to ensure the privacy of the data on the Cloud.

RESULTS (Phase 2)

Initial results from the 6 exomes revealed SNV concordance between all three pipelines was 63.5% on average. There was a far less diverse set of results when the common Cloud-based Galaxy platform was used. Although some groups—notably Group C—found large amounts of variants that only they found. The question is why? Did they use different tools or configure them differently?

CONCLUSIONS

There has been (and continues to be!) a profusion of bioinformatics tools and data processing pipelines that have been created, yet there has been a limited amount of detailed assessment of the application of these to establish best practice and specifically, recommendations and practices that ensure that they meet the rigorous requirements demanded when applied (translated) into clinical settings.

As a result of the case studies from endoVL project, we see two important aspects to reproducibility:
1. Reproducibility of software environment
2. Reproducibility of results

Eventually we see all research publications having an explicit methods section describing all aspects of the software systems and infrastructure used to create the results. This is currently not the case and the reproduction of scientific results is impacted directly.

ACKNOWLEDGEMENTS

A two-phase case study was conducted as part of the endoVL, focused on association of phenotypic databases with the variants that might be involved with cause particular kinds of DSD. The NGS data for a large family of 15 from Indonesia, 7 of whom had DSD, was made available from clinical collaborators from the Royal Children’s Hospital in Melbourne and subsequently sent to three separate bioinformatics groups across Australia for analysis.
Purpose
To develop a perimetric test algorithm that exploits spatial relationships in the visual field (VF) in an attempt to reduce test duration without detriment to precision and accuracy.

Methods
SWeLZ (Spatially Weighted Likelihoods in ZEST)

SWeLZ is a maximum likelihood Bayesian algorithm, based on ZEST. After each test presentation, the probability density function (PDF) is updated at multiple locations (Figure 1), according to a spatial model (Figure 2). Computer simulations (1000 reps) were run on 163 glaucomatous and 233 normal VFs for both ZEST and SWeLZ.

Figure 1. Example of SWeLZ update procedure. For each presentation, the location whose PDF has the largest standard deviation is tested, with the stimulus level set at the mean of the PDF. The prior PDF (top row) is updated with a likelihood function (middle row) whose shape is dependent on the observer’s response (blue: seen, red: not seen) to create a posterior PDF (bottom row). Any location in the VF that is related to that tested (columns 2 and 3) is also updated with a modified likelihood function (scaled vertically by a weighting factor and translated via an eccentricity correction factor).

Figure 2. Spatial models. A ZEST procedure using a growth pattern was used as a comparison algorithm. We tested SWeLZ with each of: 2 spatial models derived from the literature (empirical1 and geometric2) and 3 control models (nearest neighbour, random and all interconnected). The relative weights of the relationships are colour-coded according to the legend on the right.

SWeLZ does not smooth out localised defects.

To check that SWeLZ was not smoothing out localised defects, we looked at examples of typical glaucomatous VF defects (Figure 5).

Figure 5. Examples of glaucomatous VF loss. Each row is a different VF. Left column: input sensitivity; middle column: ZEST estimate; right column: SWeLZ estimate. SWeLZ and ZEST were equally able to detect: a nasal step (row 1), a paracentral scotoma (row 2), a hemifield defect (row 3) and an arcuate scotoma (row 4).

Results
SWeLZ is faster than ZEST in normal observers.

Figure 3. Global error and test times for ZEST (red) and SWeLZ (blue). Boxplots show median, IQR and 5th and 95th percentiles. Global error is the median error for each VF simulation (true – estimate). Global error is similar for SWeLZ and ZEST across spatial models (top row). SWeLZ required 20-38% less trials than ZEST for normal VFs (bottom right). Test times were similar for SWeLZ and ZEST for glaucomatous VFs (bottom left).

SWeLZ has a similar error profile to ZEST.

We looked at error split by input sensitivity (Figure 4), to investigate whether SWeLZ’s comparable error to ZEST found for whole VFs (Figure 3) translated to comparable error at individual locations. SWeLZ’s error performance was the same for all spatial maps; only the data from the empirical model is shown in Figures 4 and 5.

Figure 4. Error split by input sensitivity for SWeLZ (empirical model) and ZEST. Boxplots show median, IQR and 5th and 95th percentiles of SWeLZ error. Shaded region shows the 5th – 95th percentile range of ZEST error. Error is calculated as true sensitivity minus estimate. The 5th – 95th percentile range is similar across input sensitivities for ZEST and SWeLZ (top row). The bottom row shows the median difference in absolute error between ZEST and SWeLZ (ZEST – SWeLZ). SWeLZ and ZEST have median absolute error within 1dB of each other.

Conclusions
SWeLZ requires fewer numbers of presentations to terminate for normal VFs relative to ZEST, whilst maintaining a similar error profile. As the majority of people who sit perimetry have normal VFs, this is a significant time saving.

SWeLZ is not constrained by a growth pattern; locations can be tested in any order. With the current algorithm, the underlying spatial model makes little difference to SWeLZ’s error profile.

References
Approportion of Social Media for Knowledge Sharing By Patients With Chronic Illnesses
Nnanglefa Liika
nisika@student.unimelb.edu.au
Supervised By: Antonette Mendoza and Rachelle Bosua

Overview
This study aims to examine the appropriation of social media for knowledge sharing by patients with chronic illnesses. Social media platform is increasingly utilized by patients, leading to development of online social groups where patients share experiences and offer support to their peers on these networks. Therefore, identifying the factors that foster appropriation of these social networking sites by patients is essential. In addition, knowledge sharing behavior on this platform requires investigation to determine the behavioral factors driving willingness to contribute on these networks.

Research Aims
1. To understand how patients use social media groups to share knowledge
2. To establish an understanding of why patients participate in knowledge sharing on social media platforms
3. To understand the perceived outcomes of participation in social media groups

Research Questions
1. How do patients use social media groups to share knowledge?
2. Why do patients participate in knowledge sharing on social media platforms?
3. What are patient’s perceived outcomes from appropriation of social media groups for knowledge sharing?

Conceptual Model

References
Technology at Mealtime: Beyond the ‘Ordinary’

Hasan Shahid Ferdous | A/Prof. Frank Vetere, Dr. Bernd Ploederer, Dr. Hilary Davis
Department of Computing and Information Systems, University of Melbourne

Broad Aim
- Investigate the everyday interactions of familial uses of ICT around mealtimes.
- Explore how family members configure the dinner space and the technologies.
- Understand how technologies are used and negotiated amongst family members.
- Understand how mundane technologies support our special occasions.

Study One: Design
- Exploratory study of what technology people currently use and how they use these during family mealtimes.
- Collected qualitative data from six families through
  - two in-depth semi-structured interviews
  - home tours and
  - video recordings of two family meals.

Study Two and Three
- Broad Aims for Study two
  - Explore the possibilities enabled by sharing the content from personal devices (smartphones and tablets) during family mealtimes.
  - Developing a smartphone app that supports the sharing of content (e.g., photos, music, calendar, etc.) between devices of the family members.
  - Investigate how such system can augment the social interactions, mediate usage, and/or bring tension/constraints among the family members.

Broad Aims for Study three
- Study three works as continuation from our study one and two.
- Investigate the role shared devices can play during special and celebratory meals in families. These events are marked by special occasions (e.g., birthday, festival, etc.) and often by the presence of extended family members.
- How the contents from personal communication devices are moderated, consumed, and interacted can open interesting research opportunity.

Context and Motivation
Certain technologies like television has received much attention, especially from sociological and health perspectives. However few have explored the roles, practices, and attitudes for mobile and connected devices that now are finding a place in our everyday mealtimes. We aim to address this gap.

Research Questions
Study 1: Technology during Family Mealtime
- What are the current practices regarding technology usage during family mealtime? How technology and People are oriented together?
- What role these technologies play in supporting the commensal experience of family mealtime?

Study 2: Sharing Technology and Life
- What contents from personal devices can be shared with family during mealtime? How they are managed, negotiated, or interacted with?
- How does mobile technology use during mealtime impact on social interaction and rapport amongst family members?

Study 3: Technology at Special Mealtimes
- How extended family members impact the personal technology usage during mealtimes?

Study One: Findings
The Availability of Technology:
We base our analysis on the Heideggerian’s terms:
- Ready-to-hand technologies are available (socially and cognitively) for immediate use;
- present-at-hand technologies require a little more effort to discern their whereabouts and role in the social milieu.

Spatial Arrangement of Technology
We observed four patterns of familial arrangement around the available technologies during mealtime:
- Families arranged technologies that are ready-to-hand to enable easy access to them.
- They also arranged themselves around the technology so that all members could enjoy best.
- Various technologies were hidden but available if needed.
- Several technologies were deliberately located away from the dinner table.

Technology and Commensality
- Technology often served as a conversational resource during mealtimes.
- Certain forms of media could be backgrounded,
- Certain technologies were used as an ambiance.
- Technology was used deliberately to distract.

Conclusion
We have explored current practices around the personal and shared devices and analyzed how they support familial conversation, provide relaxation, and enable celebration, albeit not without occasional tension. Our research highlights the growing need for HCI to investigate the ever-increasing technology space in the households and the emerging and evolving family practices in this regard.

Study Three works as continuation from our study one and two. Investigate the role shared devices can play during special and celebratory meals in families. These events are marked by special occasions (e.g., birthday, festival, etc.) and often by the presence of extended family members. How the contents from personal communication devices are moderated, consumed, and interacted can open interesting research opportunity.

Fig. 1: Use of smartphone and other internet enabled device are common during mealtimes nowadays. (published with permission).
Fig. 2: Spatial orientation of the families and devices in the dining space.
Fig. 3: Father in family 1 having a discreet look at the other television (published with permission).
Fig. 4: Sharing Photos, location, and local news among family members.

References
DesTeller: A System for Destination Prediction Based on Trajectories with Privacy Protection

**Introduction**

To predict destinations of travel based on historical trajectories of the public.

**Applications:** Recommend sightseeing places; Send targeted advertisements; and Automatically set destinations in navigation systems.

**Existing solution and its limitation**

- Figure to the right illustrates historical trajectories of the public. Different colors denote different routes. The plane, shopping cart, and house icons represent different POIs. Other nodes resemble other locations.
- A user travels from $l_1$ to $l_4$. Predicted destinations are $l_7$ and $l_8$ based on the true destinations in each original trajectory to synthesize new ones.

**Proposed Solution**

Destination Prediction

Our solution is named **SubSyn** for Sub-Trajectory Synthesis. The first step is to partition and group nodes into grid cells.

Then decomposing historical trajectories into sub-trajectories and convert this information into transition probabilities using Markov model.

Computing $P=A/B$ (where $A$ is the probability of travelling from $n_i$ to $n_j$ multiplied by probabilities of all routes from $n_i$ to $n_j$, $\text{LHS in figure below}$; $B$ is probabilities of all routes from $n_i$ to $n_j$ ($\text{LHS in figure below}$).

**Privacy Protection**

- Malicious party would use SubSyn algorithm to derive sensitive location information of users, especially when they upload their locations constantly to social network websites either intentionally or unaware.
- "Exhaustive Generation": Iteratively delete each node in query trajectory and rerun SubSyn to check whether meet criteria.
- "End-Point Generation": Only need to delete end points (starting and current node). Correctness is proven in paper.

**Experimental Study**

Extensive experiments were conducted and presented in [1] using real-world dataset (i.e., trip route information from taxi company).

**Demonstration Scenarios**

DesTeller [2] is a destination prediction and privacy protection system that is interactive, user-friendly, publicly accessible, and capable of answering real-time queries. (http://spatial.cs.unimelb.edu.au/subsyndemo/)

**Conclusion**

- Proposed algorithms to perform destination prediction and privacy protection tasks robustly and efficiently.
- Built a demonstration program (DesTeller) as a dynamic webpage.
- By inputting a route, the program simulates the trip and displays predicted destinations. The behaviour of simulation can be customised.
- Both the frontend (i.e., webpage) and the backend (i.e., web service) are publicly accessible.
- The frontend can be easily adapted to other platforms such as smartphone operating systems or car navigation systems.

---


Improving Data Quality in Biomedical Trials through Automation: Feedback

Stephan Glöckner1,2; Anthony Stell1; Jemie Effendy1; Irina Bancos2; Wiebke Arlt2; Richard O. SinnoL1

1Department of Computing and Information Systems, University of Melbourne, Melbourne 3010, Victoria, Australia,
2School of Clinical and Experimental Medicine, University of Birmingham, Birmingham B15 2TT, UK

Abstract

International biomedical registries have to be driven by quality to ensure rapid translation of findings in basic science to the clinical domain.

To counter data inconsistencies, we suggest automating user feedback as an ongoing quality assurance mechanism in clinical trials.

Data Quality Feedback Categories

- Case Study – EURINE ACT Study

Background

The majority of clinical research centers are not using data quality assurance (DQA) procedures [1]. Data quality is rarely measured at the time of data collection.

The approach of clinical research centers are not using data quality assurance (DQA) procedures. It is clear that it is necessary to measure and translate findings in basic science to the biomedical and clinical domains when designing and implementing research.

Conclusion

We showed that overall data quality was improved by 35% over 6 months using quality feedback.

We propose that data quality feedback will reduce data errors in data processing and also increases the motivation of the user to contribute higher data quality.

We split feedback into three categories:

1. Real-time feedback (as data is entered into the registry)
2. Investigator feedback (periodic update reports)
3. Community feedback (e.g. league tables available to all)

Real-time Feedback

Solved: Problem: Low data entry
Solution: Motivation

Investigator Feedback

Solved: Problem: Costly data entry
Solution: Motivation

Community Feedback

Solved: Problem: Inaccurate data collection
Solution: Motivation

Proven concept

The figure shows that 59% from 838 records could be confirmed as eligible for the EURINE- ACT study.

If the records could be confirmed were data items defined as compulsory by the investigator.

Problem:

Inaccurate data collection

Costly on-site monitoring

Low data entry motivation

Community Feedback

Leaderboards of centers related to their data quality (DQS) were posted to the community.

The overall data quality was calculated based on record eligibility and data completeness and is shown in the figure. The overall data quality was 69.0.

Center Records DQS
PLWW 92.97
GRAT 116.96
ITTU 71.93
GBBI 140.89
NLEI 17.88
CRZA 56.83
GYWU 40.68
GYMU 84.61
BGSO 51.43
TOTAL 838.73

Conclusion

We showed that Overall Data Quality was improved by 13% over 6 months using quality feedback.

The efficiency of real-time feedback will be investigated in a separate case study focused on polycystic ovarian syndrome (PCOS).

Also, the tool has to show the usefulness across other domains to determine a concept of data quality dimensions (completeness, accuracy, timeliness).

References

Robust Inferences of Travel Paths from GPS Trajectories

Hengfeng Li, Lars Kulik, Kotagiri Ramamohanarao
Department of Computing and Information Systems
The University of Melbourne
henli@student.unimelb.edu.au

Problem:
How to infer a travel path on the road network from a GPS trace under noisy conditions?

Our Method:
(a) employ spatial-linear clustering to group GPS points;
(b) derive and amalgamate sub-paths of point clusters.

We propose a spatial-linear clustering algorithm to group noisy GPS points:

An example to build groups of GPS points by increasing an oriented bounding rectangle with error-bounded distance.

We derive travel paths for point clusters and combine them into a single complete path:

An example of generating possible routes bounded by the spatial-linear cluster.

We conduct a case study to analyze the impact of map matching for the estimation of traffic flow of a large area.

An example of generating possible routes bounded by the spatial-linear cluster.

Our Method:
(a) employ spatial-linear clustering to group GPS points;
(b) derive and amalgamate sub-paths of point clusters.

Method

Problem:
How to infer a travel path on the road network from a GPS trace under noisy conditions?

We propose a spatial-linear clustering algorithm to group noisy GPS points:

An example to build groups of GPS points by increasing an oriented bounding rectangle with error-bounded distance.

We derive travel paths for point clusters and combine them into a single complete path:

An example of generating possible routes bounded by the spatial-linear cluster.

We conduct a case study to analyze the impact of map matching for the estimation of traffic flow of a large area.

Related Publications

Hengfeng Li, Lars Kulik, Kotagiri Ramamohanarao: Robust Inferences of Travel Paths from GPS Trajectories. Submitted to International Journal of Geographical Information Science (under review).


Contributions

- Proposing an efficient cluster-based mapping algorithm;
- Conducting extensive experiments on both synthetic and real data sets;
- Demonstrating a case study by applying resulting travel paths for traffic flow monitoring.
Value-driven approach for Requirements Engineering in People-Oriented Software

Mohammadhossein Sherkat
Supervisors: Tim Miller, Antonette Mendoza
Department of Computing and Information Systems, University of Melbourne

Problem Statement

The challenge of requirements engineering in software design and application, such as mobile apps and social networks lies in the fact that potential users are unknown. Further, these potential users come with different kinds of personality, culture, goals, age, income and needs. Although they are not obligated to use the software, the goals of the system’s owners are attained if people use the software and its capabilities to its full potential. In this research, we call this kind of software, People Oriented Software (POS).

There are some differences between a POS and a typical business applications as stated below:

- In typical business applications, users are usually known but in POS they are unknown;
- In typical business applications, users are usually obligated to employ the software, but not in POS;
- In typical business applications, software systems support a well-defined business workflow with well-defined roles, whereas in a POS, the workflows are often ill-defined and varied.

The research question in this study is: “What is an appropriate approach for eliciting and modeling stakeholders’ requirements in the process of designing a POS?”

Proposed Approach

Data Gathering

Preparing the Owner Role Model

Qualitative Data Analysis Methods

Determining the users’ segmentations, sub segmentations and personas

Creative Problem Solving Methods

Formulating the value propositions

Preparing the users’ goals model

Preparing the users’ goals model

EQ-FAST

ESN and EQ-FAST

Modification and Integration

GFA and Indexes Designing

Designing the Channels

Owner and user
Goal Model

Value Propositions

Graphical User Interface (GUI)

Case Study

We propose to use our novel approach to analyse, design and develop a web app to help people that are homeless and those at risk of homelessness. This app will allow people find food, shelter, health and other support services. In this project, by using our approach, we aim to categorize stakeholders and elicit their emotional, quality and functional goals and needs to design an effective application. This proposed approach places an appropriate focus on the users needs and rewards to use this application for solving their problems and responding to the above mentioned needs. This project is in partnership with Crisis Care and

Legend

<table>
<thead>
<tr>
<th>Technique</th>
<th>Information</th>
<th>Evaluation</th>
<th>Maintenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESN: Emotional State Network</td>
<td>EQ: Emotional Quality</td>
<td>GFA: Goal Function Analysis</td>
<td>MADM: Multi Attribute Decision Making</td>
</tr>
</tbody>
</table>
1. Definition

- A Taxonomy for Revision Changes [1]

- Versioned Texts: text documents that have been revised and saved to another version, where the original version is directly available for comparison
- User Edits [2]
  - Factual edits: edits that alter the meaning
  - Fluency edits: edits to improve on style and readability
- Significant Changes: changes that go beyond simple string edit operations which impact meaning

Word Error Rate (WER) between two sentences, \( WER(S_O, S_R) \) takes between the value 0 to 1.

\[
WER(S_O, S_R) = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{max} (\text{length}(S_O), \text{length}(S_R))}
\]

- \( WER(S_O, S_R) = 0 \), all words in \( S_O \) aligned fully to all words in \( S_R \).
- \( WER(S_O, S_R) = 1 \), no word between \( S_O \) and \( S_R \) that can be aligned.

2. Problem Statement

Local changes are edits on sentences, where similar sentences are found between two versions. Versioned sentences, reflects local changes. Versioned sentences can be syntactically similar, but contain superficially minor differences, that nevertheless change the meaning substantially.

How do we align edits so the edits are directly comparable to form the basis for detecting significant changes between versioned sentences?

Sample of Versioned Sentences

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Revised Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stone-x-ray with current Patient Information</td>
<td>Stone-CT x-ray as Additional映射 patients with current Patient Record Information</td>
</tr>
</tbody>
</table>
| Calculate Offset of Non-_destroyed Adp. | Calculate Offset of Non-destroyed Adp.
| Select material for insert | Select material, internal anatomy, and other attributes e.g. flow profiles, extended ran of insert |

3. Structural Alignment of Versioned Sentences (SAVeS)

![Proposed Method](image)

Sample Alignment

<table>
<thead>
<tr>
<th>Single-Word Alignment</th>
<th>SAVeS Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER = 0.33</td>
<td>WER = 0.25</td>
</tr>
</tbody>
</table>

4. Results & Discussion

Discussion

- When compared to other tokenization approaches, SAVeS assists by grouping the edits as conceptual units. For example “surgeon authentication” is substituted to “authentication” or “labelled image” is substituted to “labelled annotated x-ray”.
- This form of conceptual units comparison is required before further processing to determine the significance of the edits.

5. Conclusion

- Incorporating the structure of the text affected by an edit, takes into consideration the scope of an edit in its sentential context. This may be useful as a basis for detecting significant changes between versioned sentences.
- We speculate that a phrasal representation of revisions will be better for human readability of edits.

References

Distributed stream clustering using micro-clusters on Apache Storm
Pasan Karunaratne, Shanika Karunasekera, Aaron Harwood
Department of Computing and Information Systems, University of Melbourne, Parkville, VIC 3010 Australia

Abstract
The recent need to extract real-time insights from data has driven the need for machine learning algorithms that can operate on data streams. Given the current extreme rates of data generation, these algorithms need to be able to handle data streams of very high velocity. Many current algorithms do not reach this requirement, in some cases processing only tens of messages per second.

In this work we address the problem of limited achievable throughput of stream clustering by developing scalable distributed algorithms based on the micro-clustering paradigm that run on cloud platforms.

Overview
• Maintaining summary statistics is a key aspect of stream data mining
• Micro-clusters are a widely used summary structure in stream clustering
• Real-time micro-clustering algorithm is very slow when large numbers of micro-clusters employed
• Developed distributed algorithms to run on architectures implemented on Apache Storm to improve throughput

Architectures – Common Store
• Allocators do processing independently
• Redis in-memory store used to store global state
• Aggregators form one global picture
• Our implementation allows parallelisation in aggregation step as well.

Architectures – Decentralized
• Allocators do processing independently
• Removes need for single point of synchronisation
• Trade-off with higher communication overhead

Results
• Accuracy improves with lower parallelism and higher micro-clusters
• No one architecture is consistently better in terms of accuracy
• Throughput of single machine easily eclipsed in both architectures.
• Throughput at higher parallelism values drops off in decentralized architecture, levels off in common-store

References
Summarisation Strategies for Symbolic Execution

Yude Lin
yude@students.unimelb.edu.au

Introduction

Function-level summaries [3, 1, 2] have been used to increase the efficiency of symbolic execution. However, summarisation within functions has not been attempted; the corresponding strategies are yet to be studied. We explore the use of fine-grained summaries, in which instructions become the smallest units for summarisation.

Methods

Condition pair $CP(S)$ carries the constraint information for an instruction or summarised instructions $S$:

$$CP(S) = \{EC(S), PC(S)\}$$

The entry condition $EC(S)$ is the weakest proposition under which $S$ is executed without a control flow escape. The postcondition $PC(S)$ is the strongest proposition about the state that results from execution of $S$ under its entry condition.

A path constraint is simply the concatenation of the condition pairs from all component instructions in that path:

$$x = \sin(n) > 0$$

if $(n > 0)$

$$y = x^2$$

print(x);

Figure 2: Concatenation procedure.

Summary

We explore the use of fine-grained summarisation, with which one can establish compositional symbolic execution that has stronger resistance against path explosion.

The summarisation problem is essentially: when searching for feasible paths in a tree-like structure with branch guards, how to reuse earlier results so as to reduce the cost of computing the feasibility.

Background and Challenges

Symbolic execution is a key component of modern automated software testing. It helps us understand the internal structure of the System Under Test (SUT), for a test input generation tool to produce a test suite that matches any particular program point with high likelihood. Main research directions:

1. If the program contains exponentially many paths (usually it is the case), how to mitigate the path explosion problem so that symbolic execution is less time consuming? Specifically, how to reduce the calls to the constraint solver?

2. If the program has high complexity (e.g., calling external utilities, which might not even be written in the same language as the target program), how to prevent symbolic execution from losing precision?

Preliminary Experiment Results

Figure 3: Time comparison of CSF as the depth increases. Compared to the figure are the fine-grained summarisation modes and the non-summarisation mode. Results expressed in percentages with the non-summarisation mode as 100%.

Table 1: Experiment results of running CSF with fine-grained summarisation. The number in the parentheses are the summarisation overhead. The second-last row is the average of the percentages of all tests, while the weighted average is calculated by considering the total time taken over all runs.

<table>
<thead>
<tr>
<th>Program</th>
<th>Depth</th>
<th>0.5</th>
<th>0.55</th>
<th>0.6</th>
<th>0.65</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.85</th>
<th>0.9</th>
<th>0.95</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUT</td>
<td></td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
<td>0.4</td>
<td>0.35</td>
<td>0.3</td>
</tr>
<tr>
<td>LASUM</td>
<td></td>
<td>0.9</td>
<td>0.85</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
<td>0.4</td>
</tr>
<tr>
<td>SUM</td>
<td></td>
<td>0.95</td>
<td>0.9</td>
<td>0.85</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
</tr>
<tr>
<td>SUT</td>
<td></td>
<td>0.95</td>
<td>0.9</td>
<td>0.85</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
</tr>
<tr>
<td>LASUM</td>
<td></td>
<td>0.95</td>
<td>0.9</td>
<td>0.85</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
</tr>
<tr>
<td>SUM</td>
<td></td>
<td>0.95</td>
<td>0.9</td>
<td>0.85</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
</tr>
<tr>
<td>Weighted average</td>
<td></td>
<td>0.95</td>
<td>0.9</td>
<td>0.85</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.65</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
<td>0.45</td>
</tr>
</tbody>
</table>

References

Exploring Text and Network Context for Geolocation of Social Media Users
TENSIONS IN DIGITISING BOARDGAMES

Interaction Metaphor
The metaphor holds where boardgame interfaces are extremely literal:
- Digitised cards behave like cards.
- Digitised dice are rolled to generate random numbers.
- Digitised pieces are seen to move.

The metaphor is extended where interfaces reflect a component’s function rather than its appearance or are constrained by the digital environment:
- Using action spaces or on/off lights and shadows rather than cards and pieces to indicate availability.
- Representing the score as a number rather than a position on a track.
- Selectively choosing parts of the playing area rather than the entire board, to fit the device.

The metaphor is lost where physical interaction with game components or opponents is paramount:
- Rituals of rolling dice, choosing tiles, shuffling or cutting a deck – playing with pieces (cleaning, sorting etc).
- Cheating (!).

Is a developer’s job to faithfully recreate the boardgame experience or to maximise the affordances of the digital medium?

Articulation Work
Digital games are expected to enact and enforce the rules of the game, but vary considerably in the extent to which they perform articulation tasks which support play without being a part of it.

Dividing the work & managing turn-taking:
- Turn-taking is typically managed automatically by the digital interface.
- Digital games automatically restock/refresh/shuffle discards.

Manipulating game pieces and resources:
- Typically managed by the digital interface, with an associated loss of granular control & oversight for players.

Enforcing and learning rules; supervising assigned responsibilities:
- Digital games are used as rules-learning assistants and even for rule clarification.

To players, articulation work is an essential part of the play experience; to developers, it is an unpleasant chore that is best avoided.

Informating / Theorycrafting
The digital format allows gathering & sharing of detailed data that inform gameplay.

Digitisation has allowed for new data to be gathered and compared:
- Some digital games publish information overtly; some allow played game data to be accessed.
- For many, absorbing this data and finding ways to apply it has become core to the pastime of gaming.
- It is not simply playing a game that is enjoyable, but rather attempting to solve it (“theorycrafting”).

Players memorise game statistics to increase their ability to identify the most powerful actions and to implement them at the optimal moment in the game:
- This does not necessarily improve their flexibility and skill at manipulating the game elements, but can lead to success offline as well as online.
- For these players, subjective probability is replaced with objective probability: “played badly” becomes “played incorrectly”.
- Meanwhile, other players reject the “dogmatism” of theorycrafting by rejecting a game itself instead of continuing to play organically and sub-optimally.

Sharing collected game data can fundamentally change players’ experience of the game:


Melissa Rogerson
PhD Candidate
Microsoft Research Centre for Social Natural User Interfaces
The University of Melbourne
rogersonm@student.unimelb.edu.au
1. Background

- SAT solver: Many complex NP-complete problems, like hardware verification, configuration scheduling, can be translated into satisfiability (SAT) problems for propositional logic, which can often be solved very efficiently by a SAT solver.
- SMT solver: But the low-level propositional logic has a limited expressive power while SMT solvers are decision procedures for richer theories, which can accept first order logic and reason on higher-level information and data types, such as integers and arrays. Quantifier-free bit-vector logic (QF_BV) is one such theory which we currently focus on.
- Connection between the two solvers: SAT solver is normally the back end of the SMT solver. So, if we want to improve SMT solver, improving SAT solver is one important aspect we need to think about.

2. Existing Modern Technologies for SAT Solver

- There are generally three ways to improve SAT solvers:
  1. Improve its own key elements: Three key elements of a modern SAT solver are a suitable branching heuristic, an efficient propagation, and conflict-directed learning strategy [1]. In particular, a good branching heuristic method has two properties: (a) discover conflicts as soon as possible. (b) cheap to evaluate. VSIDS variable ordering heuristic is one of the most popular methods.
  2. Propose hybrid SAT solvers: Combine SAT solver with other advanced technologies, like CP solver, and Binary Decision Diagrams (BDDs) [1] and so on.
  3. Develop parallel SAT solvers: In three approaches [2]:
     a) Split the search space and solve the subproblems in parallel.
     b) Parallelize the solving algorithm by sharing information, such as learnt clauses.
     c) Run several sequential algorithms in parallel until one result comes out by one of the algorithms.

3. Existing Modern Technologies for SMT Solver

- Bit-blasting: All modern bit-vector SMT solvers rely on bit-blasting. Bit-blasting means to encode a word-level QF_BV formula into an equivalent satisfiable Boolean representation by representing bit-vector variables as a string of Boolean variables and encoding bit-vector operations using their corresponding Boolean circuits [3].
- Researchers come up with many fancy ways to improve their bit-vector SMT solvers. There are generally three directions:
  a) Word-level simplification before bit-blasting. (level 1)
  b) Boolean-level simplification after bit-blasting. (level 2)
  c) Interact T-solver (first order logic theory) with SAT solver. (level 3)

4. What We Have Done

- Implemented a naive bit-blasting as a baseline for experiments.
- Implemented a new way of solving QF_BV formulas using advanced propagation rules on word level which is introduced in [4], with the use of the SAT solver Minisat. We don't use any bit-blasting in our solver.
- Added learning to the bitwise propagation of integers to utilise fast machine level word operations, which propagates the bit-vector on word level and explains it on bit level using the lazy clause generation concept [5].

5. Short Term Plans

- a) Do backward explanation, that is to generate explanations on demand after the propagation.
- b) Separate our propagators from original Boolean variable propagation queue. And give the original queue a priority—the original propagation of SAT solver always comes ahead of our own word-level propagation.

5. Long Term Plans

- a) Do word-level simplification on the original QF_BV formulas as pre-processing.
- b) Think about the parallel way mentioned in section 2.
- c) Think about combining CP solver with our own solver.
- d) Think about how to extend the method to solve bit vector arrays.

6. References

Optimisation Modelling for Software Developers

Authors: Kathryn Francis, Sebastian Brand, Peter Stuckey

The Problem
- They must fit
- Avoid using the overflow area

The Idea
- Don't require the developer to construct a constraint model explicitly.
- Instead ask them to answer the following two questions:
  - If you had a perfect decision maker at your disposal, how would you solve the problem?
  - If you were given a solution, how would you evaluate it?
- Once the developer has provided answers to these questions (in the form of code) we can:
  - Automatically and on-demand construct the required conventional model.
  - Provide the solution exactly as their code would have constructed it.

How it Works
1. The code written by the developer is transformed during an extra compilation phase.
2. The transformed version operates on symbolic expressions rather than concrete values.
3. At runtime the transformed version is called to obtain a problem definition in the form of a collection of expressions.
4. This definition is converted into a conventional model and sent to an external solver.
5. The program state is automatically updated to reflect the solution.
Applications

- Health & Fitness
- Medical Treatment
- Safety Circumspinction

What’s New?

Utilising user’s location context can promote the accuracy of traditional motion based activity recognition. For examples:
- An user on the highway is probably driving.
- An user moving inside the building has more probability to walk rather than drive.
- An user on the playground is probably doing sports.

Sensors

- Accelerometers
- Gyroscope
- GPS \ WiFi \ Cellular (Obtaining Locations)

Abstract Workflow

1. Re-calibrate and sampling accelerations.
2. Obtaining Geo-location and retrieving the corresponding map information.
3. Generate combined features based on acceleration waves and map information.

Signal Features

- Spectral Energy
- Fourier Transform Components
- Wavelet Coefficients

Map Information Retrieving

Indexing and Querying Techniques:
- Grid-based Indexing
- R-Tree
- Quad-Tree

Learning Methods

- Naive Bayesian
- Support Vector Machine
- Sparse Coding
- Deep Neural Network

Software Preview
Enhancing Waiting Experiences in Video Mediated Communications (VMC)

Multiactivity in VMC
Video calls are not any more just for mediating conversations any more. Nowadays, people share episodes of their everyday life with their families/friends. However, this brings about challenges of managing multiactivity.

Suspending Mediated Interactions
As a strategy for managing multiactivity, video callers put their interactant on hold to pursue other activities. This creates interstitial waiting moments for the party on hold.

Waiting Experiences in VMC
Waiting as an interstitial experience is usually linked with boredom and frustration. But surprisingly, it is not a time spent on doing nothing.

AIM
→ What role do video call technologies play in shaping waiting experiences?
→ How can we enhance such experiences?

CONTRIBUTIONS AND FUTURE WORK

Video call technologies frame waiting experiences as well as focused interactions

People’s strategies for dealing with such interstitial moments are situated in the peculiarities of these moments.

In our future work, we will explore how these experiences can be enhanced by providing local resources.

Study 1: Understanding Choreography of Waiting in VMC
Method: Observation of naturally occurring video calls

- Maintaining Availability for Re-engagement
- Scrutinising the Remote Video
- Seizing the Chance for Remote Interaction
- Engaging in Local Interactions

- Supporting Mobility & Spatial Freedom
- Supporting Awareness about the Remote End
- Supporting Opportunistic Interactions with Remote Party
- Supporting Time-filling Interactions

Study 2: Creating New Waiting Experiences
Method: Observation of semi-naturally occurring video calls

- Maintaining Audio Link Connection
- Supporting Mobility

Behnaz Rostami Yeganeh
PhD Candidate
Microsoft Research Centre for Social Natural User Interfaces, The University of Melbourne
brostami@student.unimelb.edu.au
Designing Improvised Coordination in Agent Organizations

Kathleen Keogh, Department of Computing and Information Systems, University of Melbourne
School of Engineering and Information Technology, Federation University

Abstract

Autonomous Software Agents can be used in dynamic situations where run time decision making is beneficial. BDI (beliefs, desires, intentions) agents are modelled on human behaviours and reasoning is insurable and understandable by humans and other agents. This research investigates the use of BDI agents in agent organizations to solve dynamic and complex problems such as those that emerge in Emergency Management situations. In these situations, the solution to a problem is not necessarily scripted in detail at design time. For agents to coordinate a solution, we propose a flexible meta-model and approach to facilitate agent improvisation.

Introduction

We explore meta-models for organizationally centered multi-agent system (OCMAS). We seek an organizational meta-model that facilitates agents improvisation at run time.

- share information
- coordinate behaviour
- mutually adjust individual plans
- assist others
- improvise, adopt tasks outside role

It is not possible to anticipate all situations at design time, we need to design a solution that can be flexibly applied at run time.

Main Objectives

1. dynamic coordination based on plans but with emerging details
2. What information needs to be available about the domain, the organization, the interactions and the agents? Is it provided in existing meta-models?
3. What elements of the solution can be specified at design time and what emerges at run time?
4. How can agents improvise to take on a new (un-allocated) role that needs filling when they have the capability to do so?
5. How can multiple agents coordinate together to share a role that cannot otherwise be allocated?

Methodology

Inspired by design science research [9]:

- Examine existing meta-models for Organizational Centered Multi-Agent Systems (OCMAS) [8]
- Based on limitations in existing meta-models regarding improvisation, propose new meta-model [5, 6]
- Articulate agent awareness and organizational reasoning in deliberation cycle [10]
- Describe design process for building OCMAS system [2, 7]
- Model real life problems and scenarios based on new meta-model and processes [7]

Design process

1. Define the Goal Model
2. Define the Organizational Model
3. Define the Agent Capabilities Model
4. Define the Role Model
5. Establish Social Policies to be adopted within the run-time organizational contract

Meta-Model: OJAzzIC

Organizations Juggling Adaptation, Improvising Coordination

When agents need to coordinate, an organization or ad hoc-activity is instantiated. The organization defines structure and the contract includes plans and roles. Agents deliberate aware of organizational and individual goals. Social policies guide prioritisation of goals and obligations to share knowledge and plans [3].

Conclusions

- Organizational structures provide a decompositional role of tasks that can be associated with agents to solve a problem
- It is challenging to design MAS capable of run time flexibility
- Goals broken into tasks associated with capabilities provides flexibility in terms of allocation of agents
- The ad hoc-organization defines a context for coordination of knowledge and plans
- Agents may be provided with meso level guidelines to allow improvisation using OJAzzIC organizational meta-model
- With awareness of the organizational context, agents can appropriately share information and plans

Further work

Further work is planned to verify the ability for OJAzzIC to successfully represent and implement complex problems in different domains with similar requirements.

References


Figure 1: OJAzzIC meta-model entities

- A problem is represented as a set of objectives and tasks requiring capabilities.
- Potential agent roles define capabilities and responsibilities.
- Individual agents have capabilities
- A contract [11] between agents in the organization defines:
  - role definitions
  - agent-role allocations
  - agreed shared plans
  - social policies that guide knowledge sharing obligations between agents.
- Tasks are also defined in terms of capabilities required to achieve the task [1].
- Agents as individuals have capabilities that allow matching agents to tasks outside role allocations.

Figure 2: Simulation agent system
ABSTRACT
Interactive technology at the zoo could allow close-up encounters between animals and visitors. How can we design and evaluate such technology from the perspective of animal stakeholders? How can we balance the interests of the animal with the aims of the zoo and visitors?

CONTRIBUTION
Methods for design and evaluation in the zoo context, and insights into design situated at the intersection of the natural world, animal stakeholders, public settings and social institutions. This will extend existing work in the emerging field of animal-computer interaction.

MOTIVATIONS
Animal Enrichment
Providing animals with environmental stimulation and the ability to exercise choice is essential to their wellbeing. Interactive technology such as body motion tracking could provide new and varied enrichment for zoo animals.

STUDIES
Zoo Tech: Case study
How are interactive technologies used for animal enrichment and visitor engagement? What are the interests of human and animal stakeholders? Interviews and observations focusing on three interactive systems.

Human-Animal Encounters
Zoos aim to offer opportunities for powerful connections with animals, to influence conservation attitudes and behaviours. Technology could provide a basis for profound experiences which enmesh conservation messages in animal encounters.

Kinfecting with Orang-utans
Body motion tracking using the Kinect offers the potential of a whole-body interactive experience for orang-utans and visitors. Can visitors and primates play games together? How can we design and evaluate the experiences of animals and humans?

Sarah Webber
PhD Candidate
Microsoft Research Centre for Social Natural User Interfaces
The University of Melbourne
s.webber@unimelb.edu.au

Designing Technology For
HUMAN-ANIMAL ENCOUNTERS
At The Zoo
This research will investigate how voice interaction with virtual characters affects subjective user experiences such as immersion, embodiment, identity, and control. We will apply constructivist grounded theory to analyse online discourse to form a theoretical understanding of the relationship between voice interaction and user experiences.

Interaction in virtual spaces has been dominated by the hand interacting with keyboards, mice, and game controllers. A new explosion of voice interfaces has raised the question: how does the user experience change when we use our voices to interact with virtual worlds? We will observe and interview players engaging with voice-enabled interactive games to assess the effects of social context on user experiences of voice interaction.

A fundamental contribution to the understanding of voice user interfaces is the need for design principles for voice interaction between users and virtual characters. We will propose a prototype voice interaction interface and evaluate its effectiveness through user interviews and interviews with designers.

This research will contribute to the design of voice interaction games and virtual characters, with implications for how we interact with virtual worlds.
Sequencing Operator Counts

Toby Davies  Adrian Pearce  Peter Stuckey  Nir Lipovetzky

Logic-Based Benders Decomposition

Generating GLMs is surprisingly efficient

Simple Gripper Domain

Using SAT to get better cuts

Add assumptions to SAT-planning model for each upper-bound

When UNSAT is proved, the solver identifies a subset of the assumptions responsible for failure.

Coverage

Coverage (C), Number of best bounds (+), and Dual quality scores (Q) for IPC-2011 sequential optimal track benchmarks.

Research Excellence in ICT
Wealth Creation for Australia
Transforming Situation Calculus Histories
for Efficient Reasoning

Christopher Ewin
Supervisor: Adrian Pearce
University of Melbourne

With Stavros Vassos, Sapienza Università di Roma

Introduction
• The actual sequence of actions an agent has undertaken in the situation calculus may not be ideal for reasoning tasks.
• We reorder or eliminate actions from an action sequence to obtain a modified sequence.
• The modified action sequence is semantically equivalent to the original action sequence for all generalized projection queries in the final situation.
• The modified sequence enables more efficient solutions to projection & progression problems.

Approach
• We consider two operations:
  • Reordering actions in the sequence
  • Eliminating actions from the sequence
• Applying a series of these operations produces the new action sequence.

Single Value Fluents & Resetting Actions
• Single value fluents require only one holds for any
• E.g. A particular Sokoban block can only be in one position in a given situation.
• This property is similar to functional fluents.
• We also introduce resetting actions that preserve single value fluents.
• Resetting actions affect an infinite number of but a finite number of

Results
• We develop the theoretical framework for reordering & eliminating actions in the situation calculus
• We further devise sufficient conditions for reordering & eliminating resetting actions, where the resetting actions are:
  • Context free, or
  • Have conjunctive contexts
• The table below details the average number of actions required to solve sample Sokoban problems, along with the number of actions this sequence reduces to when the eliminating and reordering operations are applied

Future Work
• Reordering of sensing actions, which can facilitate delayed progression
• Determining which (unground) actions can be reordered or eliminated off-line.
• Determining candidate actions for reordering and eliminating.
• Reordering and Eliminating actions wider classes of action theories.
• Developing action sequences which are equivalent for only particular types of projection query.
• Development of GOLOG implementation of this framework.

Conclusion
• Allows us to optimise frequently executed reasoning tasks by shortening action sequences.
• Can produce more efficient progressions of the initial knowledge base.
• Can allow non-progressable actions to be removed from an action sequence.

Acknowledgement:
NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

Contact:
Christopher Ewin
Department of Computing and Information Systems
The University of Melbourne
cjewin@student.unimelb.edu.au

table

<table>
<thead>
<tr>
<th>Number of blocks</th>
<th>Number of actions (original)</th>
<th>Number of actions (new)</th>
<th>Percentage Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>32</td>
<td>5</td>
<td>85%</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>6</td>
<td>90%</td>
</tr>
<tr>
<td>7</td>
<td>97</td>
<td>7</td>
<td>93%</td>
</tr>
<tr>
<td>8</td>
<td>130</td>
<td>8</td>
<td>94%</td>
</tr>
</tbody>
</table>

Figure 1: A solution to a single Sokoban problem
Figure 2: The new action sequence
Answer Set Programming:
Founded Bounds and Model Counting

Rehan Abdul Aziz
Supervisors: Peter Stuckey and Geoffrey Chu
University of Melbourne and NICTA

Introduction

Answer Set Programming (ASP), propositional satisfiability (SAT) and Constraint Programming (CP) are declarative programming paradigms that are frequently used for solving combinatorial problems. The thesis is divided into two parts.

Part I: Answer Set Programming for Founded Bounds

Two important features in combinatorial systems are foundedness and numeric reasoning.

Foundedness

Given a graph with three nodes a, b, c, build directed edges such that c is reachable from a.

reach_a, reach_b, reach_c, build_a, build_b, build_c

reach_a <= 1
reach_b <= 1
reach_c <= 1
build_a <= reach_a + build_a
build_b <= reach_b + build_b
build_c <= reach_c + build_c

Reachability

SAT or CP semantics considers reach_a = reach_b = reach_c = true when no edge is built. ASP, on the other hand, supports foundedness over Boolean variables such as reach variables, which gives us the right semantics.

Handling Integers

ASP and SAT solvers do not natively represent and reason about integer domains, but CP solvers do. One way to encode integer a with domain [1,100] in propositional systems is to use 100 Boolean variables a[1]...a[100] (where a[i] = true means a = i) along with the constraint that exactly one of them must be true.

Contributions Summary

<table>
<thead>
<tr>
<th>System</th>
<th>Variables</th>
<th>Foundedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT</td>
<td>Boolean</td>
<td>None</td>
</tr>
<tr>
<td>ASP</td>
<td>Boolean</td>
<td>Boolean</td>
</tr>
<tr>
<td>CP</td>
<td>Numeric</td>
<td>None</td>
</tr>
</tbody>
</table>

Foundedness over numeric variables enables reasoning about problems such as shortest paths. Shortest path rules are analogous to reachability rules. For example, as shown in the shortest path encoding above.

Part II: Answer Set Counting

Probabilistic logic programming

Consider the following problem: given a graph where each node can fail with a certain probability, calculate the probability that node x is reachable from node a. This is equal to:

\[
\# \text{ of subgraphs where } \text{reach}_x \text{ is true} \quad \frac{\text{Total number of subgraphs}}{100}
\]

This is a very hard problem, and is known to be in the complexity class \( \mathcal{A} \mathcal{P} \). Propositional semantics does not allow modelling reachability efficiently. Therefore propositional model counting, which is a well studied area, is insufficient for this.

In [3], we present an answer set counter, which implements foundedness in a state-of-the-art propositional model counter. As a result, we can efficiently solve problems like the one above.

Projected Model Counting

In [2], we describe techniques for doing model counting projected over a subset of variables. This is useful in the context when the variables whose count we are interested in, are a subset of the original variables.

Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>y</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>z</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

References


Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.
Information Security Strategies in Organisations

ABSTRACT

The University of Melbourne, Parkville, Victoria, Australia

Craig Horne, PhD student

BACKGROUND

Some recent events of global significance highlight the importance of this research: BNY Mellon by not keeping secret customer details safe? What are the customer trust and brand value implications for Bank of NY Million in costs due to an external security breach?

Craig Horne

METHOIDS

A conceptual model of ISS should be developed as a result. Three planned case studies will also contribute individually to knowledge on this concept. The literature review will be comprehensive and a publication of this is planned. A Delphi study will be condensed into an article to create supporting knowledge. A planned Delphi study of a statistically significant number of Chief Information Security Officers and Security Managers from organisations around the globe will be used to generate primary data which can then be analysed.

OBJECTIVES

1. To examine the possible relationship between the organisation’s information security strategy and the customer’s trust. 2. To assess the organisational security implications of these breaches and to determine if they are correlated. If no clear definition can be found in the literature, then a definition will be proposed. Military strategy literature will also be examined to see what can be added to support this information systems-centric concept. 3. To consider the factors that lead to different information security strategies among organisations. If you need more placeholders for titles, content or body text, just click the Bullets button on the Home tab.

CONCEPTS

Three planned case studies will also contribute individually to knowledge on this concept. A Delphi study will be condensed into an article to create supporting knowledge. A planned Delphi study of a statistically significant number of Chief Information Security Officers and Security Managers from organisations around the globe will be used to generate primary data which can then be analysed.

METHODS

A conceptual model of ISS should be developed as a result. Three planned case studies will also contribute individually to knowledge on this concept. The literature review will be comprehensive and a publication of this is planned. A Delphi study will be condensed into an article to create supporting knowledge. A planned Delphi study of a statistically significant number of Chief Information Security Officers and Security Managers from organisations around the globe will be used to generate primary data which can then be analysed.
Thank You!

An enormous amount of planning goes into the organization of an event such as the CIS DC 2015.

The CIS DC 2015 would not have been possible without the support and participation of graduate research students and staff in the Department of Computing & Information Systems.

We would especially like to acknowledge Justin Zobel, the Head of the Department of Computing & Information Systems and Rhonda Smithies, for assisting the CIS DC 2015 organizing committee with the planning and co-ordination of this event.

Our dear supervisors who have pushed us to submit papers and posters for the colloquium and have supported us in our studies thus far.

Our sponsors, Google, Microsoft, the Microsoft Research Centre for Social Natural User Interfaces, the Melbourne School of Information and the Melbourne School of Engineering, for their financial contributions to make the CIS DC 2015 possible.

Our keynote speaker Dr Anna Phan, industry partners and colleagues from other institutions.

The panelists and judges at each session, who have volunteered their time to make this event a success.

A heartfelt thank you to all our graduate research students and staff who have participated in the CIS DC 2015.

Finally, we would like to acknowledge the CIS DC 2015 DC organizing committee for volunteering their time, skills and enthusiasm and also thank the 2014 committee for their support.

The 2015 CIS DC Planning Committee!

Farzad Khodadadi
Farah Zaib Khan
Mohadesheh Ganji

Nick Downing
Yali Zhao
Yang Lei

Carlos Andres Agudelo-Serna
A special thanks to our sponsors!