Proceedings of the Seventh

Research Exposition

2010

Proceedings of the Seventh

Research Exposition

Research Expo 2010

Fredericton, New Brunswick, Canada

April 13, 2010

UNB Faculty of Compuuter Science
ISBN: 978-1-55131-140-1

Table of Contents

Research Posters

Bioinformatics

Mapping Regulatory Network from a Model Organism to a Non-Model Organism 1

Mutation Grounding Algorithm 2

Critical Infrastructures and Systems

Exploring Human Dynamics in Critical Infrastructures: An Agent-based Simulation of Stress in Hospital Surge 3

Modelling and Simulating Emergency-Response Operations 4

Hardware/software Co-Design

Acceleration of Blob Detection within Images in Hardware 5

Accelerated Decompression of Gzip in Hardware 6

Accelerating the MMD Algorithm 7

Knowledge Engineering

Text Mining & NLP based Algorithm to populate ontology with A-Box individuals and object properties 8

Managing Uncertain Knowledge on the Fuzzy Semantic Web 9

Networks and Applications

Middleware Framework for Disconnection Tolerant Mobile Application Services 10

Dynamic Admission Control for A Bandwidth Broker 11

Divisible Load Scheduling on Heterogeneous Multi-level Tree Networks 12

Investigation of channel formation in a MANET 13

FRID Network Monitor 14

Sensor Web Language for Dynamic Sensor Networks 15

Automatic Discovery of Network Applications: A Hybrid Approach 16

Security and Trust

Identifying Internet Mediated Securities Fraud: Trends and Technology 17

State of the Art in Trust and Reputation System: The Comparison Framework 18

An Online Adaptive Approach to Alert Correlation 19

IDS Alert Visualization for Network Security Monitoring and Analysis 20

Simulating Intrusion in NS-2 21

Spatial Data Structures

I/O-efficient Rectangular Segment Search 22

A Data Structure for Efficient Search of Objects Moving on a Graph 23

I/O-Efficient Spatial Data Structures: Observations on the d-Dimensional Grid File 24

2009 Research Publications 25

2009 PhD Theses 39

2009 MCS Theses 41

Mapping Regulatory Network from a Model Organism to a Non-Model Organism

Rachita Sharma, Patricia Evans, Virendra Bhavsar

Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada E3B 5A3

Introduction

Determination of regulatory networks from available data is one of the major challenges in bioinformatics research. A regulatory network of an organism is represented by a set of genes and their regulatory relationships, which indicate how a gene or a group of genes affect (inhibit or activate) production of other gene products. Some organisms such as yeast, *Arabidopsis thaliana* and fruit fly have been investigated very thoroughly by biologists as model organisms, being simpler and having shorter life cycles. We have developed a system to map the regulatory network from a model organism (source genome) to a non-model organism (target genome), about which less information is known.

Objectives

- Map regulatory elements and their relationships (links) from a model organism to a non-model organism
- Compare different methods used to map the regulatory links

Regulatory Elements Mapping

- Map transcription factors based on (Figure1)
  - sequence similarity - **TFbl**
  - protein family classification - **TFf**
  - protein sub-family classification - **TFsf**
- Map target genes based on (Figure 2)
  - sequence similarity - **TGbl**
  - transcription factor binding site (TFBS) motifs - **TGbs**
  - sequence similarity and TFBS motifs - **TGblbs**

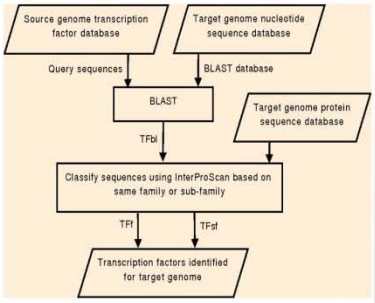


Figure 1: Method to map transcription factors from a source genome to a target genome

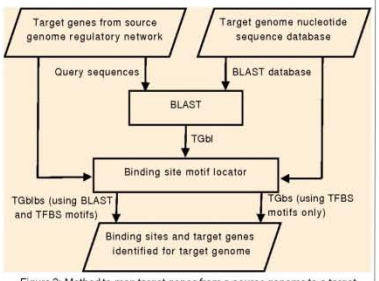
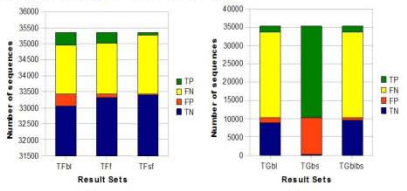


Figure 2: Method to map target genes from a source genome to a target genome

We have used *Saccharomyces cerevisiae* as the source genome and *Arabidopsis thaliana* as the target genome for experimentation in this work. We evaluated the mapped transcription factors (TF) and target genes (TG) by comparing them to the available transcription factor data and binding site data of *Arabidopsis thaliana*, respectively. The result sets are compared as shown in Figure 3(a) and 3(b) based on True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). We found that transcription factor mapping based on same protein family classification (TFf) has better performance than the other two result sets based on sequence similarity (TFbl) and both sequence similarity and same protein sub-family classification (TFsf). Target genes set predicted using TFBS motifs only (TGbs) is the best result compared to the other result sets based on sequence similarity only (TGbl) and sequence similarity and TFBS motifs both (TGblbs). Most of the target genes have been determined using the TFBS motifs only.



(a) (b)

Figure 3: Comparing result sets of (a) transcription factor mapping methods and (b) target gene mapping methods

Regulatory Elements Integration

We integrate the mapped regulatory elements (TF and TG) to predict regulatory links for the target genome as shown in Figure 4.

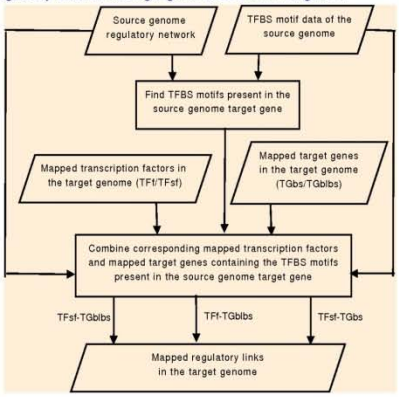


Figure 4: Method to integrate mapped regulatory elements into regulatory links for the target genome

Regulatory Links Verification

We set rules to evaluate the predicted regulatory links using gene expression experiments. The expression values of both transcription factor and target gene from a regulatory link should be present in the experiment to evaluate that regulatory link. An experiment can either Confirm (C), Contradict ($\bar{C}$) or be Neutral (N) for any regulatory link as shown in Table 1.

Transcription Factor Expression	Gene Expression	Result
Yes	Yes	C
Yes	No	$\bar{C}$
No	Yes	N
No	No	--

(a)

Transcription Factor Expression	Gene Expression	Result
Yes	Yes	$\bar{C}$
Yes	No	C
No	Yes	C
No	No	$\bar{C}$

(b)

Table 1: Rules to verify predicted regulatory links with type (a) positive gene regulation and (b) negative gene regulation

The Confirmed value (c) for a regulatory link represents the number of experiments that support that link. The experiments contradicting the regulatory link are part of the Contradicted value ($\bar{c}$). The rest of the experiments that neither support nor contradict the regulatory link, but do provide additional information about the regulatory link, are represented by the Neutral value (n).

The results for the three predicted regulatory links sets are shown in Table 2. Rows 5 to 9 in the table show the different conditions used to evaluate the results based on Confirmed, Contradicted and Neutral values. The preferred set TGbs for target gene mapping does not give the best results for the regulatory elements integration step because the additional predicted target genes in this set contribute to a lot of false regulatory links in the target genome. This set finds most of the target genes but not the correct target genes corresponding to the right transcription factor.

	TFsf-TGblbs	TFf-TGblbs	TFsf-TGbs
Number of mapped transcription factors	118	434	118
Number of mapped target genes	2252	2252	35181
Number of links mapped	43423	480524	536154
Number of regulatory links analyzed	3056	8621	6085
$c = 2\bar{c} + n$	2090	2628	2995
$c = 3\bar{c}$	2109	2375	2968
$2\bar{c} = c < 3\bar{c}$	58	344	30
$\bar{c} = c < 2\bar{c}$	103	215	100
$c < \bar{c}$	786	5687	2987

Table 2: Regulatory links confirmed for *Arabidopsis thaliana* using 43 gene expression experiments

Table 2 shows that the set TFsf-TGblbs of predicted regulatory links has better results than the other two sets, based on having a significantly higher proportion of regulatory links that are confirmed by the gene expression experiments. Therefore, integrating the mapped TFs based on protein sub-family classification along with the mapped TGs based on sequence similarity and TFBS motifs produces the best results for regulatory links. These results indicate that the regulatory relationships are conserved between genomes and can be mapped from one genome to another.

For future work, more information about gene regulation at different stages of gene expression can be incorporated, along with the new data that becomes available for the non-model organism, to map the regulatory network.

Publications:

Sharma, R.; Evans, P.A.; Bhavsar, V.C., "Transcription Factor mapping between Bacteria Genomes", *International Journal of Functional Informatics and Personalised Medicine*, **2009**, Vol. 2, 4, 424-441.

Sharma, R.; Evans, P.A.; Bhavsar, V. C., "Mapping Regulatory Network from a Model to a Non-Model Organism", Submitted to *ACM International Conference on Bioinformatics and Computational Biology* (August 2-4, 2010).

Canadian Potato Genome Project · BIO Atlantech · Atlantic Canada Opportunities Agency · New Nouveau Brunswick Canada · Natural Sciences and Engineering Research Council of Canada · Conseil de recherches en sciences naturelles et en génie du Canada · Canada · GenomeCanada

Mutation Grounding Algorithm

Jonas B. Laurila[1], Rajaraman Kanagasabai[2] and Christopher J. O. Baker[1]

[1]University of New Brunswick, Saint John. [2]Institute for Infocomm Research, Singapore.

April 13th, 2010

Motivation

Protein mutations are derived from in vitro experimental analysis and their impacts described in detail in scientific papers. Reuse of mutation impact annotations is an important subfield of bioinformatics for which mutation grounding is a critical step.

We present a method for grounding of textual mentions from scientific papers describing mutational changes made to proteins. We distinguish between grounding of mutation entities to database entries and positionally correct grounding on amino acid sequences extracted from protein databases.

Conclusion

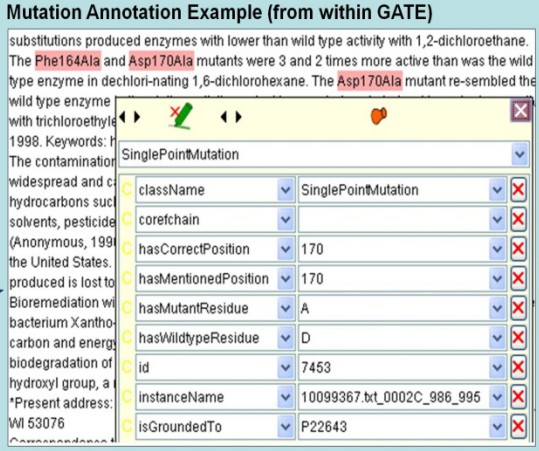
Automated reuse of mutation impact information from documents is now an achievable milestone, given the respectable performance of our grounding algorithm.

In combination with mutation impact extraction from sentences, the mutation grounding algorithm will facilitate the construction of unique datasets suitable as training material for predicting the impacts of genomic variations and the extraction of genotype-phenotype relations.

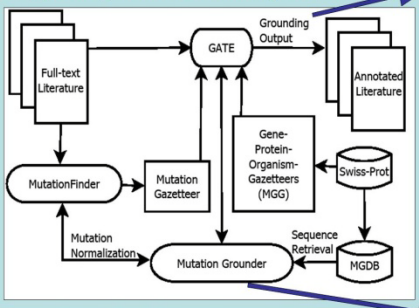
Evaluation

To evaluate the method for mutation grounding a gold standard corpus was built using the COS MIC database. Three target proteins/genes were considered, PIK3CA, FGFR3 and MEN1. Full-text papers containing more than one single point mutation and only about one single gene were chosen, with a total number of 63 documents.

Mutation Annotation Example (from within GATE)



Entity Recognition & Grounding Framework



Grounding Workflow

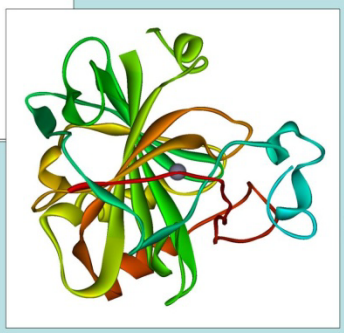
1. Retrieve **protein** and **gene** mentions.
2. Retrieve all related **accession numbers** from *MGDB*, discard all but the most occuring.
3. Retrieve all **organism** mentions and discard accession numbers not related to retrieved organisms.
4. Retrieve all unique **mutation** mentions, normalize with *MutationFinder* and try to **fit as many as possible onto the sequences** corresponding to the accession numbers still left.
5. The accession number and corresponding sequence on to which most mutations are grounded is now considered as the correct one for the entire document.

Grounding Algorithm

```
1:  function GROUND(M, sequence)
2:      GSet ← ∅
3:      for all m_i ∈ M do
4:          for all m_j ∈ M \ {m_i} do
5:              regex ← BUILDREGEX(m_i, m_j)
6:              while MATCH(regex, sequence) do
7:                  displacement ← CALCULATEDISPLACEMENT(regex, sequence)
8:                  G ← {m_i, m_j}
9:                  for all m_k ∈ M \ G do
10:                     if sequence[m_k.position − displacement]=m_k.wildtype then
11:                         G ← G ∪ {m_k}
12:                     end if
13:                 end for
14:                 GSet ← GSet ∪ {G}
15:             end while
16:         end for
17:     end for
18:     GSet ← MOSTMUTATIONS(GSet)
19:     return LEASTDISPLACEMENT(GSet)
20: end function
```

Performance

Corpus	Precision	Recall	Corpus size
PIK3CA	0.86	0.70	30
FGFR3	0.89	0.66	26
MEN1	0.54	0.32	7
Average	0.84	0.64	63



Acknowledgements

-New Brunswick Innovation Foundation

-NSERC Discovery Grant awards to Christopher J. O. Baker

Exploring Human Dynamics in Critical Infrastructures:

An Agent-based Simulation of Stress in Hospital Surge

Alexis Morris,
Faculty of Computer Science, UNB Fredericton

Motivation:

Is it possible to predict systemic failures in critical infrastructures that are due to human factors?

Critical infrastructures are complex, socio-technical systems that involve a coordinated mix of people, technologies, and organizations. These ensembles have a unified objective; to protect important assets, predict and prevent losses, and promote a stable operation of a system. Such systems comprise a social and a technical subsystem which are interdependent and inter-organizational networks of human-to-human, human-to-technology, connections. These are directed at a high level by policy makers, decision-makers, and translate to decisions of groups, and their use of technologies.

The problem is that such systems are prone to failure due to indirect consequences of decisions, flaws in organizational policy, and a poor understanding of the complexity of the holistic inter-dependencies within and among organizations.

The way of improving is to understand both social and technological aspects that cause global system failures. The technological systems are typically well defined, and well researched problems. The human problems, however, are less clearly understood. Hence the research question**: How can one use the techniques of simulation, both agent based, and system dynamic, in order to predict failures due to human dynamics and organizational policies?**

Methodology:

- Understanding how to represent non-discrete, and "fuzzy concepts" in simulations of human factors.
- Defining a taxonomy of important human factors in critical situations, eg. stress, trust, culture, personality, and emotion.
- Defining a particular case study and base simulation environment.
- Enhancing agents with interesting human factor models.
- Analyzing behaviours over time and in different contexts.
- Modelling policies and important facets of several critical infrastructures (hospitals, police stations, fire stations, etc).
- Assessment of results and policies that predict system failure.
- Understanding the benefits of holistic simulation using top-down, and bottom up methods (namely Agent-based, and System Dynamics techniques).

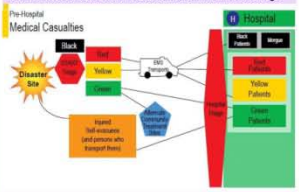
Experiment 1: Stress in Hospital Surge

Medical support systems are a first responder in crisis situations. As such, it is common that these systems face massive overloads when the number of casualties is high and they are near a disaster site. Hospital surge capacity is often a function of the number of beds available, incoming patients, current patients, and staff available. We want to model the psychological factor of stress on these individuals in different configurations, with a focus on victim arrivals, nurses, doctors, technicians, and resources. It should be possible to concretize this notion, and measure it against surge capacity.

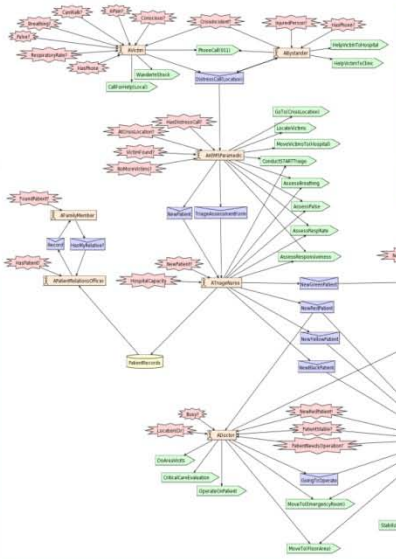
We intend to model this scenario using NASA's Brahms agent platform. Stress modeling makes use of Hobfoll's Conservation of resources as a starting point. This will eventually be explored as a system dynamics model, of rates, levels, and feedback loops.

Hospital Surge Scenario:

First responders in a crisis are trained in categorizing casualty illnesses through a process known as triage. Below triage sorts victims into four classes for eventual treatment and discharge.



From: "Surge, Sort, Support: Disaster Behavioural Health for Healthcare Professionals Textbook" DEEPCenter, University of Miami. 2006



Agent Design: System Overview

The basic agent based simulation design is shown here, minus an appropriate stress model.

Agents, their percepts, activities, and message communication are shown.

After implementation the numbers of victims and different configurations of hospital workers will be instantiated, allowing for understanding surge capacity, patient distress, and worker burnout.

Discussion:

Medical informatics is an immediately relevant domain for studying human dynamics in critical systems. Results should show that failures due to stress, and burnout can be measured, and predicted. This could lead to describing policies and better configurations of work environments. Future work will involve the merger of this hospital simulation scenario with others in an inter-dependent critical infrastructure.

Adaptive Risk Management Laboratory

Prof. Mihaela Ulieru
Canada Research Chair
Director ARM Laboratory

UNIVERSITY OF NEW BRUNSWICK
UNB

Modelling and Simulating Emergency-Response Operations

William Ross
Faculty of Computer Science, University of New Brunswick

Motivation

- ➢ Disastrous incidents over the past decade (e.g., Hurricane Katrina) have exposed serious weaknesses in the emergency-response capabilities of modern countries.
- ➢ We are investigating a subset of the factors negatively impacting emergency response; specifically, we are interested in minimizing the effect of inter-organizational conflict to improve response effectiveness.
- ➢ At present, we are exploring the normative (i.e., policies) and structural dimensions of the various organizations involved to investigate how conflicts can be minimized.

Scenario

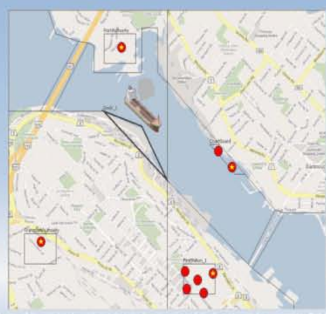


Figure 1. A view of the simulated world, including the burning ship and the four simulated organizations (Port Authority, Firefighters, Coast Guard, and Transport Authority (bottom)) are denoted by stars. In this world, there are three possible outcomes: (i) the fire may be stopped early, while it is still small; (ii) the fire may escalate to engulf a larger portion of the ship, but still be stopped before the fire spreads further; and (iii) the fire may cause an explosion onboard the ship, causing fire to spill out onto the dock. The effectiveness of the response—based on the ability to contain the incident—is a function of the interaction among the involved organizations (i.e., the effect of their combined policies).

Methodology

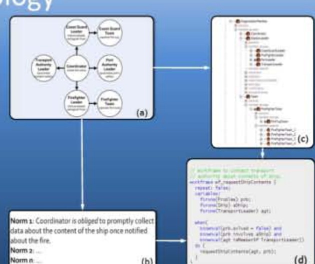


Figure 2. Our general methodology for building accurate representations of our organizations involves capturing their structure, as well as their policies (or norms). We begin by modelling these dimensions in OperA, a language used to model agent organizations. We then take these models (the structural model (a) and normative script (b)) and implement them in Brahms, an agent-simulation language. The structural model is converted into Brahms groups (c), while the normative script is translated into Brahms workframes (d). These workframes specify the condition (in the "when" clause) in which group members may follow the indicated policy.

Experiment

Table 1. The four rule sets used to test the three variables of interest in our experiment.

Variable	Rule Set 1	Rule Set 2	Rule Set 3	Rule Set 4
1. Coordinator promptly collects data from Transport Authority (Normative Dimension)	¬Obligated	Obligated	¬Obligated	¬Obligated
2. Fire tug team enters exclusion zone (Normative Dimension)	Prohibited	Prohibited	¬Prohibited	Prohibited
3. Fire tug team owned by (Structural Dimension)	Coast Guard	Coast Guard	Coast Guard	Firefighters

Table 2. Aspects in the simulation which are randomized to account for environmental uncertainties.

Aspect	Reason
1. Communication time	Some messages take longer to convey than others; people are not always immediately available
2. Team effectiveness	Teams have different levels of fatigue and experience
3. Travel times	The time of day, traffic levels, and condition of roads/water impact travelling time
4. Explosion threshold	The explosion is not a strict function of the response: it may take longer (or shorter) for the explosion to occur even when the response is identical
5. Weather conditions	Temperature, wind direction, and wind strength affect the fire

Results

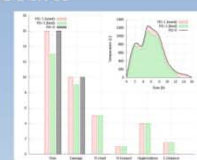


Figure 3. The results produced by Rule Set 1 are close to the no-response case (black). Because the coordinator does not quickly collect data about the ship, the fire grows more rapidly than expected, an exclusion zone is established as a result, and the fire tug team is unable to assist at the scene when requested (policy conflict).

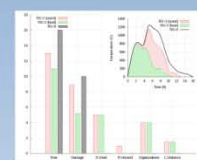


Figure 4. The results produced by Rule Set 2 show a large difference between the best and worst cases (i.e., low reliability). This is because the fire tug team sometimes arrives at the scene before (and sometimes after) the exclusion zone is established due to the randomized aspects presented in Table 2.

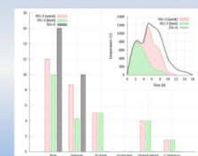


Figure 5. The results produced by Rule Set 3 again show a large difference between the best and worst cases. Because the coordinator lacks important data, sometimes the fire tug team is requested before the explosion, and sometimes after the explosion.

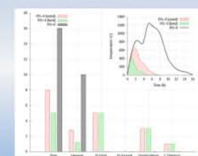


Figure 6. The results produced by Rule Set 4 show that in all runs of the simulation the explosion of the ship is prevented. Having one less organization saves valuable communication time. In our experiment, this rule set produced the most effective response.

Conclusion

- ➢ We constructed organizational representations for the four organizations involved in our scenario.
- ➢ We then examined the impact of two normative dimensions and one structural dimension on the simulated response.
- ➢ Finally, using a set of original metrics capable of showing the effectiveness (i.e., the best case) and reliability (i.e., the difference between the best and worst cases) of the response under various configurations, we determined that Rule Set 4 produced the best response in our experiment.

NOTE: Our experimentation approach can be used to test other variables and variable combinations in our simulation.

Adaptive
Risk Management
Laboratory

Dr. Mihaela Ulieru
Canada Research Chair
Director ARM Laboratory

Dr. Nicola Bicocchi
Postdoctoral Fellow
Project Manager

UNIVERSITY OF NEW BRUNSWICK UNB

Acceleration of Blob Detection within Images in Hardware

Alexander Bochem / Kenneth B. Kent
University of New Brunswick
Faculty of Computer Science
v8w2q@unb.ca ken@unb.ca

Outline

• Implementation and evaluation of image processing algorithms on FPGA architecture

• Parallelization of image processing algorithms

• Experimental implementation of blob detection methods for evaluation purposes

• Computation of blob center point by bounding box and center of mass

• Comparison of performance and precision with similar solutions on General Purpose Processor architectures

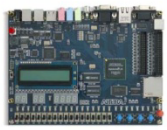
Motivation

Invention of a passive tracking system for estimation of position and orientation of the user in virtual environments

Representation of the user by laser emitting device, projecting a unique pattern to allow 6 degrees of freedom

Material

Altera DE2 board with Cyclone II



Quartus II Web Edition (ver. 9.0)
- Implementation in SystemVerilog
- Simulation with Waveform files
- Compilation and programming

Analog video input source providing simplified image material

Problem

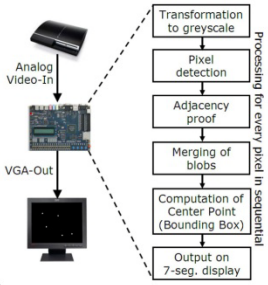
Computer Vision tasks are required in many applications. We need to find a balance between high image resolution and fast processing speed. Most of the time either one of those needs to be shortened to fulfill the desired criteria.

The problem addressed in this project is the detection of binary large objects (blobs) in a continuous video stream and compute their center point. It is related to a project at the Bonn-Rhein-Sieg University of Applied Sciences in Germany where the invention of a multi user interaction device for 3D projection environments is ongoing.

By employing standard image processing algorithms on FPGA architectures we are working to achieve a performance gain for real-time interaction.

Solution

The system performs a transformation from color to greyscale. All pixels are processed sequentially and tracked as part of a blob if their brightness is above a defined thres-hold value. Selected pixels are checked for adjacency to already detected blob pixels in the current frame. Adjacent blobs are merged and their center point is computed by the bounding box method which is displayed on the seven-segment-display of the DE2 board.



Results

For evaluating the performance the very same blob detection method has been implemented on a GPP architecture. It scanned the frame pixels sequentially and computed the blobs center points by bounding box measurement.

It could be shown that the FPGA approach can perform with at least 20 frames per second faster than the very same program logic on a GPP architecture. This also includes the additional video processing on the FPGA system to convert from analog to digital before performing the blob detection while the GPP system could directly grab whole frames from the digital video file.

The extension of the system by a center of mass based computation of the center point is going to be one of the next steps in the future.

The design of the immersion square requires the computation of three video streams in parallel, since all three sides of the cube need to be tracked. This and the parallelization aspect of the FPGA architecture for the processing of the single image frames have not been taken into account until now.

Accelerated Decompression of Gzip in Hardware

Joe Libby / Kenneth B. Kent
University of New Brunswick
Faculty of Computer Science
g6x2d@unb.ca ken@unb.ca

Outline

- Implementation of Zlib Inflate() compression algorithm on a Field Programmable Gate Arrary (FPGA)
- Acceleration of Zlib Inflate() by exploiting parallelism and pipelining
- Hardware specific optimizations of Zlib Inflate() algorithm

Motivation

Video game terminals require fast responsiveness in order to hold the attention of the user. Users will remain at a terminal longer if there is little to no perceived delay between activating a feature and seeing the feature on screen.
The goal of this project is to minimize load times by maximizing the throughput of image loading. This will be realized by offloading image decompression to a hardware module running on an FPGA.

Background

GZip
•Lossless compression algorithm
•Based on LZ77 Compression Algorithm
•Replaces data with references to matching data that has already passed through the encoder/decoder

Example

How the compression works

Input Stream: Blah blah blah blah blah!

The encoder will iterate through the stream finding the first repetition of lah b which produces:
Blah b[D=5,L=5]

But the encoder can increase the amount of compression realized as it moves through the rest of the stream
Blah b[D=5, L=18]!

Target Platform

•The target platform for this system is a general purpose computer utilizing an onboard FPGA and Flash memory for long term data storage.

Problem

Decompression in software on a general purpose CPU much too slow to provide an adequate end-user experience. Software decompression struggles to meet even the data transfer rate from Flash storage.

Solution

Offload the decompression processing from the general purpose CPU to an FPGA.

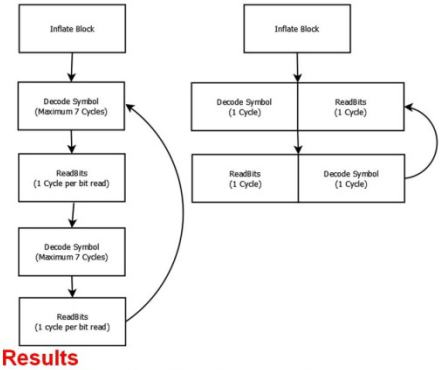
Challenge
•Identifying why the software performs poorly on a general purpose CPU and finding a hardware solution that would alleviate this problem

•First problem: Gzip library is large.
  •Some components and code can be removed and still leave a functional decompression core

  •Solution: Tiny Inflate: Open source C implementation of the decompression routine Inflate() only.
    •Tiny Inflate also solved the problem of dynamic memory management by using a static model.
    •While better for this project, Tiny Inflate still contains code that is wasteful and must be optimized for hardware

•Second Problem: Optimizing Tiny Inflate

  •The following diagram shows how the main decompression cycle was optimized using parallelism and pipelining.
  •Left flow shows original and the right flow shows the optimized version.



Results

The current version of the decompression core is performing at a maximum decompression rate of 20 Megabytes per second. In comparison, on a 2.8 Ghz Pentium 4, the same benchmarks had a maximum throughput of 11 Megabytes per second.

UNIVERSITY OF NEW BRUNSWICK
Reconfigurable Computing Research Group
FACULTY OF COMPUTER SCIENCE

Accelerating the MMD Algorithm

Michael Schlösser and Kenneth B. Kent
University of New Brunswick
Faculty of Computer Science
m09j8@unb.ca ken@unb.ca

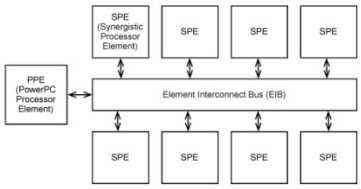
Motivation

The MMD algorithm is used to synthesize and optimize reversible logic networks. The runtime of the algorithm is strongly dependent on the size of the input network and can result in a runtime of several hours up to several days. Due to the time complexity, the algorithm uses heuristics in order to compute near optimal solutions.
The goal of this project is the acceleration of the algorithm. An accelerated version could be used to compute bigger inputs more efficiently or for more exact results. To achieve this goal the algorithm will be parallelized using both the Cell Broadband Engine (Cell BE) and an OpenCL based GPU implementation.

Materials

The Cell Broadband Engine (Cell BE)

The Cell Broadband Engine (Cell BE) is an implementation of the Cell Broadband Engine Architecture (CBEA), defined by IBM. It is based upon the 64-Bit PowerPC Architecture and extends it in order to provide a processor specification for parallel computing. For this project a Cell BE processor inside a Playstation 3 (PS3) is used.



The processor consists of one PowerPC Processor Element (PPE) which serves as program logic controller and several Synergistic Processor Elements (SPE) optimized for data-rich operations.

Graphics Processing Unit (GPU)

GPUs are highly parallel many-core processors commonly used in order to calculate real-time computer graphics. In the last couple of years they have been newly-discovered and utilized for computationally intensive general purpose tasks. Due to their massively parallel architecture they have the potential to outperform any common CPU currently available on the market.

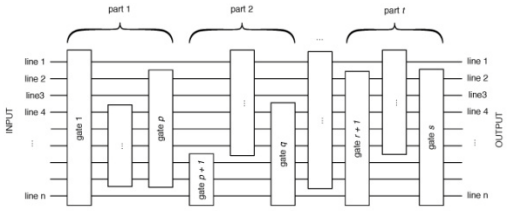
OpenCL

The Open Computing Language (OpenCL) is an open standard for hardware device agnostic parallelization of software applications. The standard is maintained by the Khronos Group which also is responsible for standards like OpenGL.
Any hardware architecture that is able to provide an implementation of the OpenCL specification is able to run OpenCL programs. The concept of writing a parallel program once and run it on heterogeneous devices like CPUs, GPUs or the Cell BE make OpenCL interesting for parallelization tasks.

Method

Both the Cell BE and GPUs are highly parallel hardware architectures. In order to be able to accelerate the MMD algorithm, goal of this research work is to exploit the underlying hardware as good as possible. This involves taking several hardware specific characteristics into account. These include e.g. number of parallel processing units or memory bandwidth and size, just to name a few.

The contemplated approach focuses on the aforementioned optimization step, the so called template matching. That means an already synthesized input network will be loaded and divided into *n* independent parts. The MMD algorithm will then be executed on every single part in parallel. In the end all *n* results will be collected and merged into a new and optimized reversible network.



Anticipated Results

With respect to the project goal, an acceleration of the algorithm can be expected. Especially for big input networks the parallel approach has several advantages that should lead to a shorter runtime of the algorithm. The current sequential approach only takes a small part of the whole circuit into account. That means only a small sequence of gates get optimized although the algorithm could work on several parts simultaneously. The parallel nature of this approach makes it possible to target the idling parts of the circuit and therefore deliver faster result.

References

D.M. Miller, D. Maslov, and G.W. Dueck. *A transformation based algorithm for reversible logic synthesis.* In Design Automation Conference, 2003. Proceedings, pages 318–323, June 2003.

IBM. *Cell Broadband Engine Architecture.* https://www-01.ibm.com/chips/techlib/techlib.nsf/ techdocs/1AEEE1270EA2776387257060006E61BA/$file/CBEA_v1.02_11Oct2007_pub.pdf (2009-18-12), October 2007.

Khronos OpenCL Working Group. *The OpenCL Specification.* http://www.khronos.org/registry/ cl/specs/opencl-1.0.48.pdf, June 2009.

REVLIB - *The Online Resource for Reversible Functions and Circuits.* Website: http://revlib.org/.

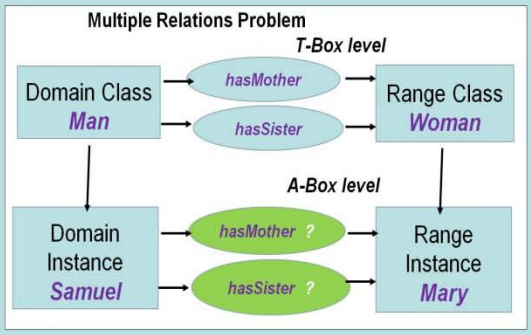
Text Mining & NLP based Algorithm to populate ontology with A-Box individuals and object properties

Alexandre Kouznetsov and Christopher J. O. Baker,
University of New Brunswick, Saint John, joint work with **Innovatia Inc**

April 13th, 2010

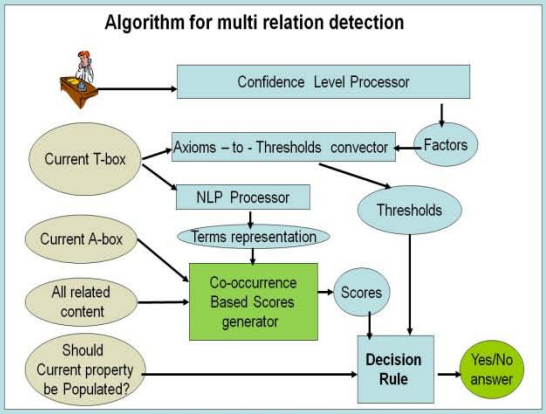
Motivation

Ontologies can play a very important role in information systems, particularly in facilitating information retrieval and data integration. In this contribution we present a semi-automatic method for extracting information, specifically named entities and their relations, from texts and populating a domain ontology. While previous work has proposed solutions to extract named entities and populate them to classes an ontology, we are focused on the problem of accurately extracting and populating multiple relations between the same named entities and presenting them as distinct object properties between A-box individuals in an OWL-DL ontology.



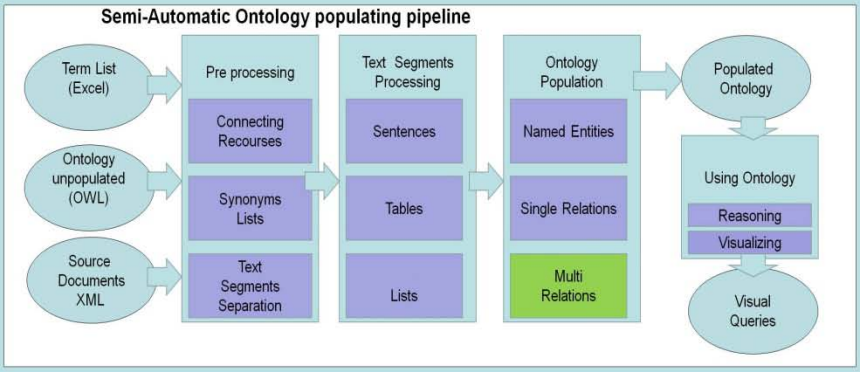
Methodology

Ontology-based information retrieval applies Natural Language processing (NLP) to link text segments, named entities and relations between named entities to existing ontologies. In our algorithm we leverage a customized gazetteer list, including lists specific to object property synonyms and score A-box property candidates by using functions of distance between co-occurred terms. Using ontology reasoning we build Confidence Thresholds on A-box property candidate scores. A-box Property prediction and population based on these scores and thresholds.



Algorithm main modules

1)NLP processor: *to extract term(s) to represent each object property*
2) Confidence Level processor: *to convert settings to Threshold Factors*
3) Axioms-to- Thresholds convector: *to extract A-box related axioms and convert into decision boundary Thresholds on property candidate scores*
4) Co-occurrence Based Scores generator: *to calculate scores based on normalized distances between domain, range and property terms*
5) Decision Rule *to populate properties that obtained scores over Thresholds*



Implementation tools

Java, OWLAPI, GATE/JAPE, PELLET

Delivery

Semi- Automatical Ontology populating pipeline prototype is under testing on BioMed (Lipids) and Telecom (Innovatia/Nortel) ontologies

Acknowledgment

We would like to thank Bradley Shoebottom for his help with Telecom knowledge engineering.

Managing Uncertain Knowledge on the Fuzzy Semantic Web

Jidi Zhao, Harold Boley, and Weichang Du

Motivation

※**Limitations of Precise Reasoning**
- ※concepts without well-defined boundaries often have to be defined with 'artificial' boundaries
- ※originally uncertain relationships have to be forced into precise relationships for knowledge representation
- ※distorting reality and expert thinking
- ※giving up important properties
- ※loss of authentic representation

※**Uncertainty Reasoning**
- ※uncertainty is an intrinsic feature of real-world knowledge
- ※based on known uncertain facts (evidence)
- ※applying uncertain axioms and rules
- ※resulting in conclusions that are uncertain to some degree
- ※better resembling human reasoning in its use of approximate information and uncertainty to generate decisions

Solution

※**Fuzzy Description Logic fALCHIN**
- ※a fuzzy extension to the Description Logic ALCHIN
- ※based on Vague Sets
- ※ **fALCHIN** includes **fuzzy** concepts, roles, and constructors

※**Fuzzy Knowledge Base**
- ※fuzzy axioms and fuzzy assertions

※**Core Reasoning Algorithm**
- ※based on tableau algorithm with fuzzy extension

※**Various Inference Services and Procedures**

※**F-OWL (Fuzzy OWL)**
- ※a fuzzy extension to OWL 1 & 2
- ※abstract concrete syntax / functional-style syntax
- ※core semantics based on fALCHIN

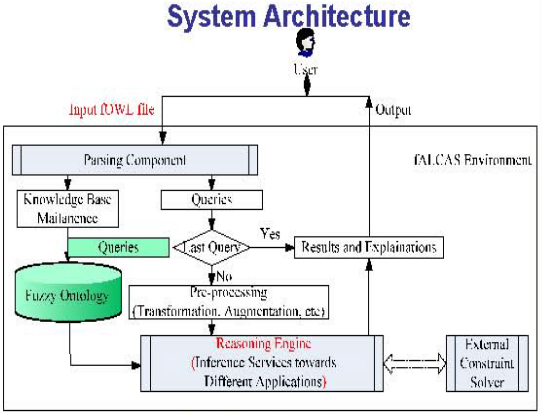
※**Prototype Implemented in Prolog: fALCAS**

Description Logics and OWL

※**Description Logics:**
- ※logic-based knowledge representation formalisms
- ※about the conceptual knowledge of arbitrary domains
- ※DLs basics include
  concepts, roles, individuals, constructors, axioms and assertions

※**OWL : Web Ontology Language**
- ※W3C's OWL 1 & 2 recommendations for the Semantic Web
- ※based on Description Logics
- ※three OWL 1 species: OWL Lite, OWL DL, and OWL Full
- ※three OWL 2 profiles: OWL 2 EL, OWL 2 QL, and OWL 2 RL

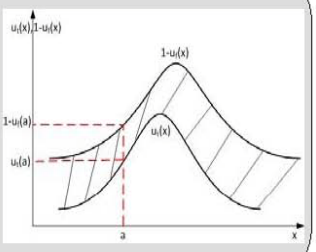
System Architecture



Fuzzy Logic and Vague Sets

※ **Fuzzy Logic:**
- ※ membership function u(x) with single value (D->[0,1])
- ※ no accuracy measurement

※ **Vague Sets:**
- ※ interval-valued
- ※ $[u_t(x), 1-u_f(x)]$
- ※ truth-membership function : $u_t(x)$
- ※ false-membership function: $u_f(x)$
- ※ positive and negative evidence
- ※ accuracy measurement



Application Services

Medical Application Scenarios

※Consistency Checking (general)

※Fuzzy Instance Entailment (patient eligibility)
- ※instance role entailment
- ※instance concept entailment

※Fuzzy Concept Subsumption and Similarity (symptom and diagnosis comparison)

※Fuzzy Retrieval (patient documents)
- ※top-k instances retrieval
- ※threshold- θ instances retrieval

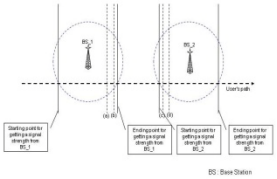
Middleware Framework for Disconnection Tolerant Mobile Application Services

Sangwhan Cha, Weichang Du, .Bernd J. Kurz

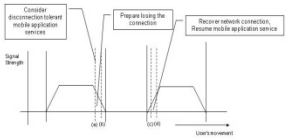
Introduction

Mobile services are prone to failures caused by the disruption of an active wireless access network connection due to the user's movement to other networks or signal blocking (shadowing).
Thus, proper mechanisms for disconnection tolerant mobile application services are needed. we propose a middleware framework that transparently performs required functionality for users in order to provide continuous mobile services in case of network disruption.
Such middleware framework provides an effective disconnection tolerant mobile application service.

Research Problem

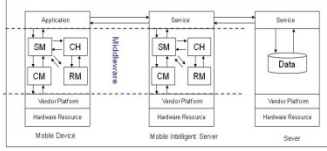


Figures shows a network disruption between base station 1 and base station 2 when a mobile device moves from point (a) to point (d). Thus, mobile multimedia application services cannot be provided properly during network disconnection from point (b) to point (c).



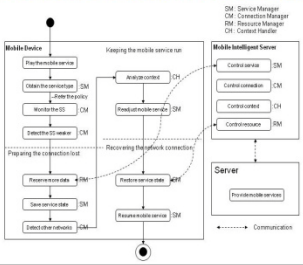
Appropriate mechanisms are needed for preparing a persistent and resumable mobile service before the current wireless access network is lost, making sure that a mobile application continues to run on the mobile device, until reconnection occurs through another detected wireless network, and recovering the execution of the mobile service after the wireless access network reconnection.

Middleware Architecture Overview

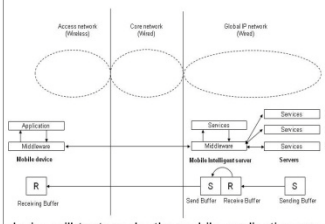


SM : Service Manager
CH : Context Handler
CM : Connection Manager
RM : Resource Manager

Middleware Framework Design



Solution



MIS is located in the fixed network and has reliable connections to the application server.

When the disruption in the access network occurs, the middleware on the mobile device will try to make the mobile application run as long as possible with limited resources until reconnection,

Real-World Applications

✓ Current Existing Applications

- Stored multimedia
  : Mobile VOD, Mobile IPTV
- Real time Multimedia
  : Mobile Video Call
  : Mobile Conference
  : Mobile streaming multimedia
  (Watch child care, Watch house ,…etc.)
- Others
  : Mobile Game,
  : VR (Virtual Reality) application
  : Etc.

✓ Future Applications

- Real time Multimedia
  : Mobile Cloud computing
- Virtual Reality (VR)
  : Mobile Training
  : Mobile Education
  : Mobile Medical Application
  : Mobile E-Commerce
  : Mobile Entertainment
  : Mobile Manufacturing

✓ **For more information, Refer to the paper :**
S. Cha, W. Du, and B. Kurz, "Middleware framework for disconnection tolerant mobile application services" in proceedings of the 10th communication networks and services research conference (CNSR 2010), Montreal, May, 2010.

Communication Networks and Services Research (CNSR)

This research is supported and funded through CNSR by Bell / Aliant and ACOA by an AIF research contract.

Dynamic Admission Control For A Bandwidth Broker

Chenguang Gao, John DeDourek, Przemyslaw Pochec
Faculty of Computer Science, University of New Brunswick Fredericton, NB, Canada

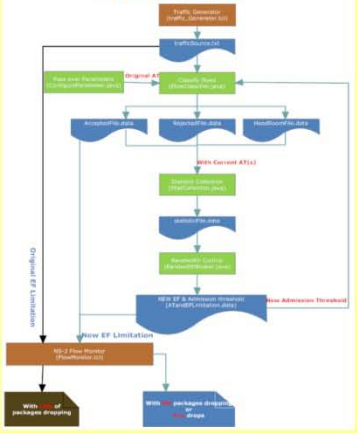
INTRODUCTION

- Multimedia and real-time applications require high quality services.
- Quality of Service (QoS) provides better service such as reduction of the number of dropped packets, delay, jitter, and out-of-order delivery.
- The IETF proposed the Differentiated Services (DiffServ) , which classifies flows with DiffServ code point (DSCP) and the Per-Hop behavior (PHB).
- A Bandwidth Broker (BB) manages the resources based on the Service Level Agreement (SLA) by controlling the network load and by accepting or rejecting bandwidth requests.

OBJECTIVE

- Design a scheme to provide dynamic bandwidth management in a DiffServ domain with a bandwidth broker.
- Simulate the proposed scheme in NS-2 and analyze results with respect to performance of the admitted streams, and with respect to the cost of unused reserved resources

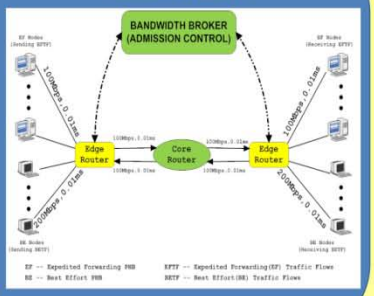
Overall Flowchart



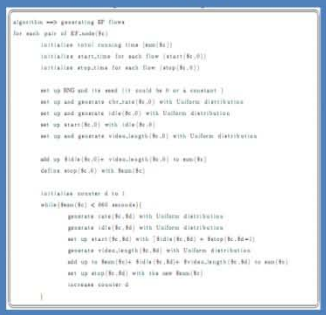
Methodology

- Flow Generator - randomly generate EF flows and store them into a file
- NS2 DiffServ Script – provide the DiffServ, monitor packet dropping , read incoming flow file and bandwidth file
- Smart Admission Control - read flow files, inspect network load, predict and generate future threshold for EF traffic

Network Topology



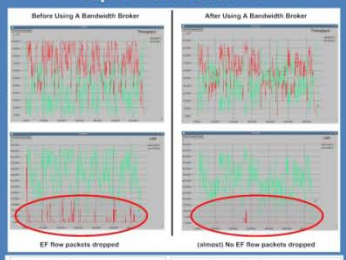
Flow Generating Algorithm



Admission Control Algorithms

- Static Admission Control:
  $AT(0) = AT(1) = AT(2) = \ldots\ldots = AT(N) = \text{Initial AT}$
- Dynamic Admission Control: $AT(N) = AT(N-1)_{(optimal)}$
  $\eta = \alpha \times \text{UsedCapacity} - \beta \times \text{UnusedCapacity} - \gamma \times \text{RejectFlows}$ $(\alpha = \beta = \gamma = 1)$

Experiments and Results



Conclusion

- The simulation shows improved QoS for the EF traffic with dynamic admission control (very few EF packets dropped).
- Performance measured with the metric η is higher for the dynamic algorithm than for the static algorithm.
- The proposed scheme successfully provides dynamic bandwidth management with a bandwidth broker.

Chenguang Gao, Graduate Research Expo, UNB,2010

Divisible Load Scheduling on Heterogeneous Multi-level Tree Networks

Mark Lord (g1835@unb.ca), Eric Aubanel (aubanel@unb.ca)

University of New Brunswick Faculty of Computer Science

Problem Statement

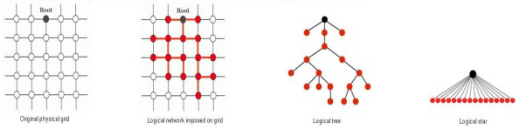
Divisible Load Theory (DLT) is a framework used for modeling large data-intensive computational problems and is applied using Divisible Load Scheduling (DLS) algorithms. There is an inherent parallelism to be exploited in large computational loads (large data files to be processed) that can be partitioned into smaller pieces to be computed concurrently. The granularity of these loads is assumed to be infinite, allowing each fraction of load to be computed independently. Some example applications that take advantage of this divisibility property include: matrix computations, Kalman filtering, database searching, image and signal processing, etc. [1]. Finally, the goal of DLS is to determine the schedule in which to distribute an entire divisible load among the processors to be computed in the shortest time possible.

A heuristic known as ITERLP2 [2], based on ITERLP [1], can calculate a DLS with result collection for a Star (Single-level tree) network with latency costs in polynomial time. ITERLP2 is solves $O(n^2)$ linear programs (minimization problems to minimize DLS solution time) as opposed to OPT (Optimal solution) that requires solving $O((n!)^2)$ linear programs.

The creation of divisible load schedules on heterogeneous multi-level processor trees is still an open problem [1], so it is the goal of this research to create a heuristic for calculating a DLS with result collection for a Multi-level tree networks with latency costs using a similar approach to ITERLP2 in a recursive fashion. It is expected that solution times produced by a DLS for a logical multi-level tree will be lower compared to an equivalent logical Star network with the same underlying physical network. The number of linear programs solved will decrease at the cost of an increase in the size and complexity of linear programs. This decrease in the number of linear programs solved greatly improves the time to calculate a DLS.

Mapping Logical Networks in Physical Networks

Purpose: Given a random physical network consisting of processor nodes of fixed heterogeneous Computation (E), Communication (C), and Latency (L) parameters, find both Star (single-level tree) and Multi-level tree logical networks which are imposed/mapped to processor nodes and network links in the physical network.



Example of finding logical Star and Multi-level tree structures which have been imposed on a random physical network (note: processors/links used in either mapping need not be the same).

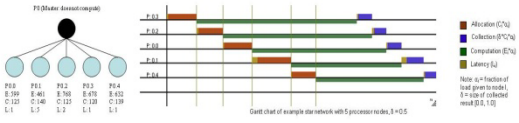
Building Logical Star: Starting from source (root) to destination, define a single logical network link with communication cost (C) by choosing highest cost link in the original physical network from source to destination. Latency cost (L) is found by adding the latency cost of each physical link and assigning the sum to the logical link. Computation cost of each physical processing node is preserved.

Building Logical Tree: All physical network link costs and processing node costs are preserved.

DLS on Star (Single-level tree) Networks

Background: A simple network model for DLS is the star network with a master-worker platform. It consists of a master processor and a set of worker processors connected via the points on the star. One assumption to be made for this network model is that only one communication between the master and any given worker processor can happen at a time, i.e., the single-port model. The master processor does not do any computation and has the sole responsibility of adhering to the divisible load schedule performing allocation and collection [1].

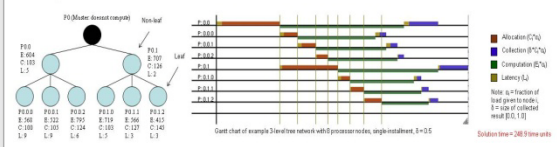
Given an arbitrary divisible load schedule, the DLS process starts with the orders of allocation/collection and associated load fractions being provided to the master processor. This master processor accepts a divisible load for the first phase of DLS. This phase begins with the master processor dividing and allocating load fractions to each worker processor in the order of allocation provided by the divisible load schedule. One-by-one the worker processors receive their load fractions until the transfer is complete. Upon each successful transfer, the master allocates the next load fraction to its associated worker processor. Since the load fractions can be computed independently, the worker processors begin computing their allocated loads upon receipt. Finally, when all load fractions have been allocated, the master begins collecting back the results produced from the worker processors to then form a complete solution [1]. The overall result from DLS is that each load fraction can be calculated in parallel.



Example logical star network mapped to a random physical network with 5 nodes having parameter ranges E = [400, 800], C = [100, 150], L = [1, 5].

Multi-level Tree Heuristic I (Single-installment)

Background: Multi-level trees are far more complex to determine a schedule on than star networks. One would need to find the equivalent processors for each non-leaf node from the bottom up and then sorted at each level. One assumption to make with these multi-level tree networks is that every non-leaf node (excluding the root) have the capability to compute and communicate at the same time. This is done by delegating the communications to a front-end (communications processor) while simultaneously computing a load fraction [3]. Under this assumption, the mathematical model used for a star network (without front-ends) does not apply. The results of DLS with result collection on multi-level trees is expected outperform those of a single-level tree (star network) due to the master processor being able to delegate the communication costs to the non-leaf processors (with front-ends) to allow load fractions to be further broken up to be distributed concurrently.



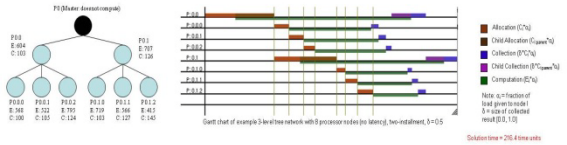
Example logical multi-level tree network mapped to a random physical network with 8 nodes having parameter ranges E = [400, 800], C = [100, 150], L = [1, 5].

The heuristic algorithm solves for a set of equivalent processors and associated load fractions at each level from bottom up. These equivalent processors represent all sub-trees/processors connected at each level in the tree, allowing a load to be split multiple times down the tree. This involves solving $O(n_i^2)$ linear programs for each non-leaf node i, where n_i is the number of nodes below node i. It is expected that the total number of linear programs solved is less than the number required to solve for a DLS on a star network.

Algorithm details omitted.

Multi-level Tree Heuristic II (Two-installment)

Background: Similar to Multi-level single-installment, the two installment approach uses the notion of front-end processors. The major difference is that every non-leaf node will be allocated its own load fraction as the first installment for which they can begin computing immediately after receiving. Upon beginning to compute its own load fraction, the second installment of load that is to be computed by the children and children's children is allocated to the associated non-leaf node. This in turn allows the network to begin computing earlier than the single-installment approach and is expected to produce lower solution times.



Example logical multi-level tree network mapped to a random physical network with 8 nodes having parameter ranges E = [400, 800], C = [100, 150], L = [1, 5].

The heuristic algorithm solves for a set of equivalent processors in a similar fashion to single-installment by taking the bottom-up approach. Again, this involves solving $O(n_i^2)$ linear programs for each non-leaf node i, except the complexity of each linear program is increased slightly. Again, it is expected that the total number of linear programs solved is less than the number required to solve for a DLS on a star network.

The existing implementation of the two-installment DLS doesn't yet support latency costs, but is currently being worked on.

Algorithm details omitted.

Conclusions and Remaining Work

In this work we have described the design for a heuristic algorithms to calculate a DLS for a Multi-level Tree network, with either single-installment and two-installment, in polynomial time.

Remaining work on this project includes the completion of implementation of two-installment heuristic considering latency costs, completion of algorithm to impose logical star and multi-level tree networks onto random physical networks, and testing to ensure both correctness and efficiency.

References

[1] H. Watanabe O. Beaumont A. Ghatpande. H. Nakazato, Divisible load scheduling with result collection on heterogeneous systems, Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium (2008), 1-8.

[2] M. Lord. Divisible Load Scheduling with respect to Communication, Computation, and Latency Parameters. (Bachelor's honour's thesis, UNB, Summer, 2008).

[6] V. Bharadwaj, D. Ghose, V. Mani, T.G. Robertazzi. Scheduling Divisible Loads in Parallel and Distributed Systems. IEEE Comp. Society, 1996.

Faculty of Computer Science UNB

NSERC CRSNG

Investigation of channel formation in a MANET

Kerul Patel, John DeDourek, Przemyslaw Pochec
Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada

Introduction

• An ad-hoc network is a collection of wireless mobile nodes forming a temporary network without the aid of any stand-alone infrastructure or centralized administration. In an ad-hoc network, the nodes not only act as hosts but also assist in establishing connection by acting as routers that route data packets to/from other nodes in the network

• In a Mobile Ad-hoc Network (MANET), the nodes have a tendency to move and so the topology may change dynamically and unpredictably. Due to this spontaneous and dynamic nature, routing in MANETs is proving to be an interesting and challenging problem for researchers. To overcome the routing problem, various routing protocols have been developed

• We investigate the formation of a channel in a MANET taking into account parameters such as range of a wireless node, area size, number of nodes in the topology, and routing protocol used.

Simulation Parameters

• Network Simulator 2 (NS-2) is used to investigate the scenario.

• NS-2 supports various MANET routing protocols and enables a user to create and modify the wireless topology, number of nodes and their mobility, data traffic, and various other parameters.

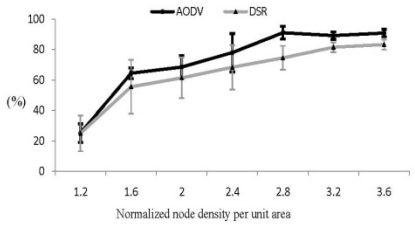
Topology size (A)	1250 x 1250 m
Simulation time	50 sec
Range of wireless node (R)	250 m
Number of nodes (N)	30, 40, 50, 60, 70, 80, 90
Routing protocols used	AODV, DSR

Performance Metrics

• We introduce a parameter *normalized node density* per unit area. The unit area is defined as the area covered by the transmission range of a wireless node. The normalized node density can be calculated using the formula R*N/A.

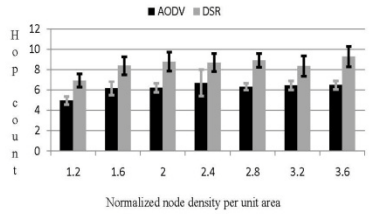
1. Percentage of packets received

The ratio of the number of packets successfully received at the destination to the total number of packets sent from the source. This is the main performance metric as it shows how successfully the connectivity between the two nodes in a network is maintained. 100% packets received indicates that the connectivity was available all the time.



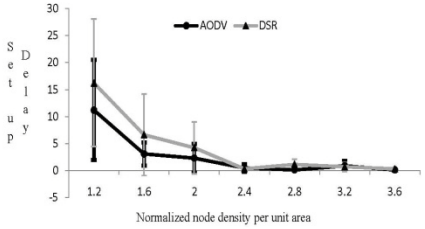
2. Hop count

The number of times a packet was forwarded by wireless nodes before reaching its destination.



3. Set-up delay

The time from when the first packet is sent until the first packet reaches its destination i.e. time taken to form a channel for the first time.



FRID Network Monitor

Xi Ruan, Wei Lu, Andrew Merrithew, Hanli Ren, Ali Ghorbani

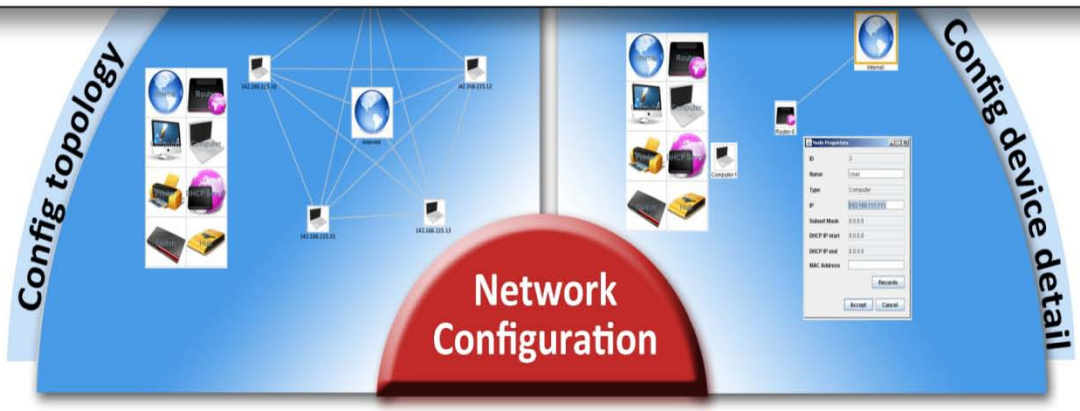
Information Security Centre of Excellence

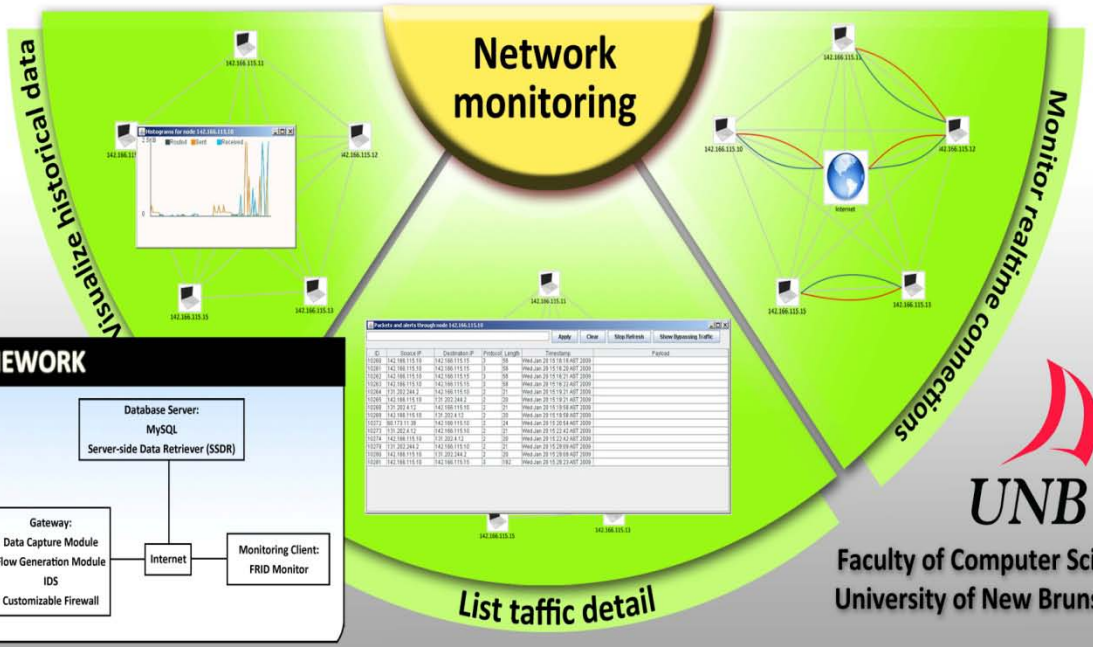
iSCX

INSTRUCTION

FRID Network Monitor system provides mutiple functions which aim to help a network administrator to monitor the network security status.

- Network topology configuration
- Network activities visualization
  - Visualize realtime and historical connection activities
  - Visualize botnet activities
  - Visualize network security alerts
- Generate security reports





FRAMEWORK

UNB
Faculty of Computer Science
University of New Brunswick

Sensor Web Language for Dynamic Sensor Networks

Gunita Saini and Bradford G. Nickerson
University of New Brunswick, Faculty of Computer Science

Motivation:

To automatically display changes in Wireless Sensor Network (WSN) to the web.

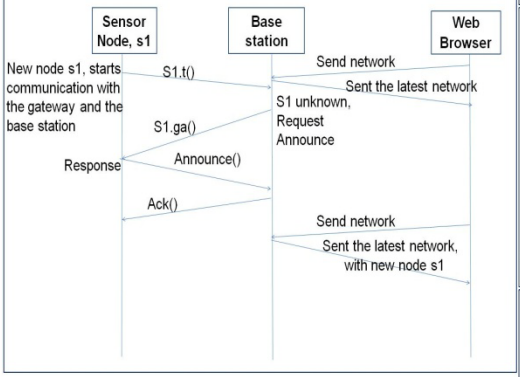
Objectives:

- Current Wireless Sensor Network can be seen and queried from any browser.
- Perform lab tests, by turning sensor nodes on and off.

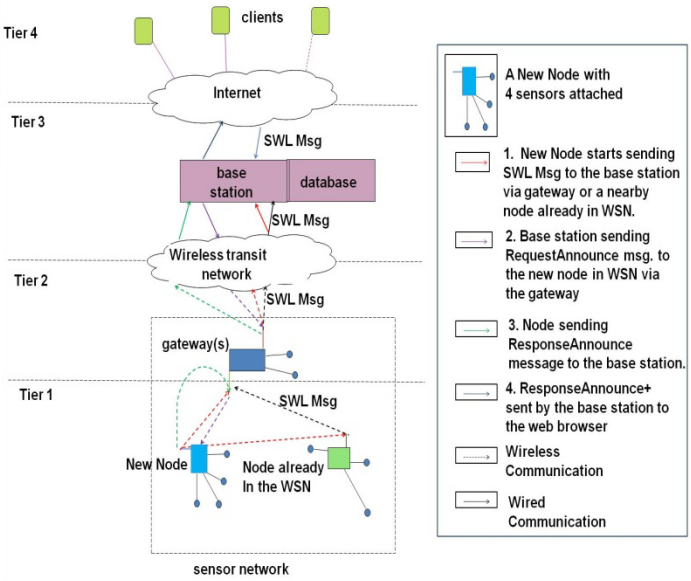
A portion of Adaptive Sensor Web Language (SWL) grammar

...
15. *RequestAnnounce* ::= {*Identifier* .ga();}
16. *Response* ::= *ResponseConstructor* | *ResponseArrayCon* | *ResponseArrayValue* | *ResponseConfig* | *ResponseAnnounce*
17. *ReponseConstructor* ::= {*Identifier.ldRest* ([*Param*]);}
18. *ResponseArrayCon* ::= {*ArrayRef.ldRest* ([*Param*]);}
19. *ResponseValue* ::= {*Identifier.ldRest* = *Exp* ;}
20. *ResponseConfig* ::= {*Configuration* {{*varDecl* }* {*Exp* }*}}
21. *ResponseAnnounce* ::= {nid={*Hexdigit*}+16;lt={N|S}*deg,min,sec* ;s={*IntegerType*};ln={E|W } *deg,min,sec* ;n={*IntegerType*}(;SID={*Hexdigit* +})} | {*NodeConfig* }
22. *NodeConfig* ::= {*Configuration*{{*varDecl*}*{*Exp* }*}}
...
49. *deg* ::= {*Digit* }+2
50. *min* ::= {*Digit* }+2
51. *sec* ::= {*Digit*}+5
...

Communication between the web browser, base station and a new sensor node in Wireless Sensor Network (WSN), sequence diagram



Sensor Web Components – Dynamic Communication



ResponseAnnounce message payload - 22 bytes, total size of the ResponseAnnounce message is 28 bytes sent from the sensor node to the base station.

NodeUUID	
Node Latitude	Session ID
Node Longitude	n
Sensor UUID1	
...	
Sensor UUIDn	

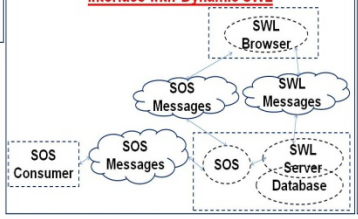
UUID is the Universally unique identifier, unique throughout the SWL. n is the total number of sensors attached to the node. Message size of ResponseAnnounce message is 3+n.

Sample format of ResponseAnnounce message

```
response{
nid=46f52dc76c6042319;
lt=N45,94,599999;s=97;
ln=W66,63,339128;n=3;
SID=c6cf046dc1fa4e638;
SID=cfdf449efd554fe0b;
SID=391e816927484bcbb;
}
```

Sent from the new node in WSN and collected by the base station.

Proposed Sensor Observation Service (SOS) interface with Dynamic SWL



Sample SWL Browser screen

Triangle represents a sensor node, circles represent the sensor(s) attached to the node and rectangle represents the gateway.

Sponsored By:
Faculty of Computer Science UNB
MITACS
FI NB IF INNOVATION

Automatic Discovery of Network Applications: A Hybrid Approach

Mahbod Tavallaee, Wei Lu and Ali A. Ghorbani
Information Security Centre of Excellence, Faculty of Computer Science

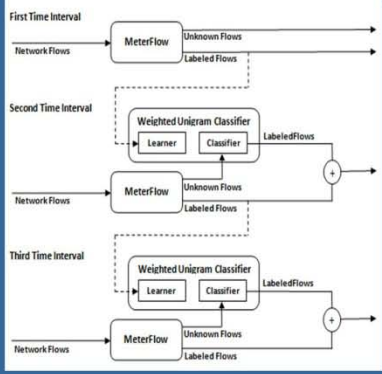
Abstract:

Automatic discovery of network applications is a very challenging task which has received a lot of attentions due to its importance in many areas such as network security, QoS provisioning, and network management. In this poster, we propose an online hybrid mechanism for the classification of network flows, in which we employ a signature-based classifier in the first level, and then using the weighted unigram model we improve the performance of the system by labeling the unknown portion. Our evaluation on two real networks shows between 5% and 9% performance improvement applying the genetic algorithm based scheme to find the appropriate weights for the unigram model.

Contributions:

- Employing the unigram of packet payloads to extract the characteristics of network flows. This way similar packets can be identified using the frequencies of distinct ASCII characters in the payload.
- Applying a machine learning algorithm to classify flows into their corresponding application groups.
- Assigning different weights to the payload bytes to calculate the unigram distribution. We believe that depending on the applications those bytes that include signatures, should be given higher weights.
- Applying genetic algorithm to find the optimal weights that results in improvement in the accuracy of the proposed method.

Hybrid traffic classification scheme:



Unigram distribution of an HTTP flow:

GET / HTTP/1.1\r\nHOST: www.google.com\r\nConnection: keep-alive\r\nUs

Letter	Freq.	Letter	Freq.	Letter	Freq.	Letter	Freq.
G	1	E	1	T	4	space	4
/	2	H	2	P	1	1	2
3	3	\r	3	\n	3	O	1
S	1	:	2	w	3	g	2
o	5	i	2	e	5	c	2
m	1	C	1	n	3	t	1
l	2	k	1	p	1	-	1
a	1	v	1	U	1	s	1

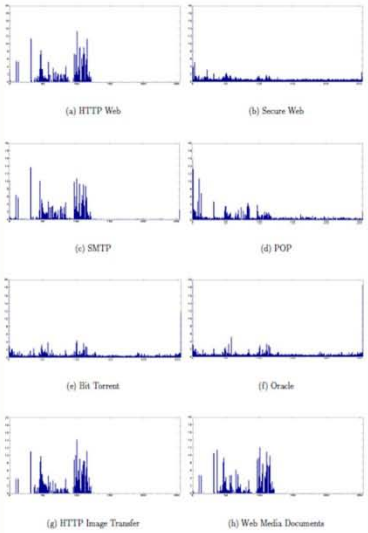
(a) First 64 bytes of the source payload

HTTP/1.1 200 OK\r\nDate: Fri, 18 Dec 2009 17:31:23 GMT\r\nExpires: -

Letter	Freq.	Letter	Freq.	Letter	Freq.	Letter	Freq.
H	1	T	3	P	1	/	1
space	9	2	3	0	4	O	1
K	1	\r	2	\n	2	D	2
a	1	t	1	e	3	:	4
F	1	r	2	i	2	,	1
8	1	c	1	9	1	7	1
3	2	G	1	M	1	E	1
x	1	p	1	s	1	-	1

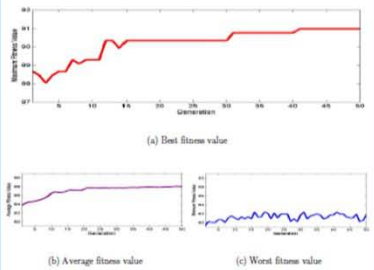
(b) First 64 bytes of the destination payload

Unigram distribution of network applications:

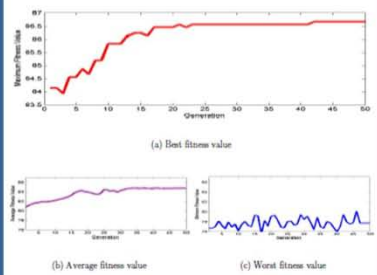


Applying Genetic Algorithm to find the optimal weights:

1.First data set (ISCX Network)



2.Second data set (ISP Network)



Results:

	ISCX Network	ISP Network
Base Accuracy	81.93%	81.72%
First Generation Accuracy	83.69%	80.74
Optimal Accuracy	90.97%	86.55%

Applications	Total Flows	Unknown Flows	Successfully Classified	Total Accuracy
MSN Messenger	53	0	0	100%
Bit Torrent	103	27	18	91.26%
HTTP Web	121	12	6	95.04%
SMTP	74	4	2	97.30%
Windows File Sharing	91	8	5	96.70%
Total	442	51	31	95.48%

Identifying Internet Mediated Securities Fraud: Trends and Technology

Jake van der Laan, Brodie M. Shannon, and Christopher J.O. Baker
University of New Brunswick and New Brunswick Securities Commission

The Problem

The world wide web makes it much easier to commit securities fraud. Securities regulators do not become aware of these frauds until after the damage has been done. Identifying these operations early is difficult using traditional surveillance methods. Innovative technology driven solutions are required. Creating such tools requires both a multi-disciplinary understanding and approach to this problem.

What is Securities Fraud?

Securities fraud takes many different forms, but most involve either the selling of investments or the manipulation of their value. In many cases, the internet is used to advertise, promote, or effect the actual securities transaction.

Prevalence on the Web

Since 2001 internet mediated securities fraud has been one of the most common types of securities fraud and will likely become the dominant form in the coming years.

Types of Internet Securities Fraud

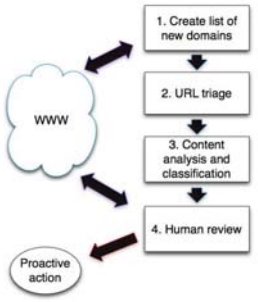
- Illegal distributions (sometimes referred to as boiler rooms. Securities offered by unregistered individuals or without proper disclosure.)
- Market manipulation (artificially drive a particular stock price up or down, in order to reap a profit from this change.)

Drivers of Internet Fraud

The internet makes identity easier to falsify and more difficult to authenticate, lowers the economic resources needed to set up, expands the "target market", and makes the proceeds of crime easier to divert. All of these factors weigh against effective law enforcement.

Scamalyzr

In 2009 we developed "Scamalyzr", a simple word based text classification tool which searches a corpus of continuously updated new website instances (retrieved from the web) for prevalence of a pre-determined set of relevant words, and then ranks them based on the presence and frequency of these words.
The tool is currently being used by the New Brunswick Securities Commission and identifies potential securities fraud websites on a daily basis.



Scamalyzr process flow

Results

Between March and September 2009, Scamalyzr processed over 13 million domains and classified these into 4 categories: suspect, not suspect, no-content, and parked. The no-content and parked categories were regularly re-scanned for content and re-classified accordingly. Fom the list of suspects, the highest ranking domains were reviewed by an analyst as a result of which boiler rooms were identified, and were placed on the New Brunswick Securities Commission website's Caution List or referred to law enforcement agencies and banks in the United States, the United Kingdom, Australia, and Canada.

Opportunity

Even though the internet has enabled boiler rooms, it has also created a potential way of fighting them which was not previously available. This potential lies in the fact that the desired information (the promotional website and details of the touted investment) is accessible, as soon as the website is launched. The problem, of course, is finding it.

Challenges and Further Work

Despite providing valuable information, the system currently generates a high incidence of false positives ($> 80\%$). A more rigorous information extraction process will be implemented in the next phase of Scamalyzr development, using an ontology and machine learning approach (OBIE). We intend to extract information consistent with the concepts and relationships defined in the domain ontology, and then to maintain that information as instances of OWL ontologies, in turn enabling access to the output of the OBIE system by the Semantic Web.

The absence of useful metrics in assessing our current system is a problem and our future work will adapt established "precision" and "recall" metrics for information retrieval to our context.

Fraudulent investment website detected by Scamalyzr

The Importance of Web Science

It is truly surprising how people continue to surrender their hard earned money to someone they have never met. It is difficult to reconcile how people develop a high degree of trust in someone who calls them on the phone and convinces them, over a period of a few days by pointing them to a website and without actually meeting face to face, to invest in something which in most cases does not even exist.

Dealing with this is very much a multi-disciplinary problem. It requires not only an understanding of the technologies being used by scammers, but also, and perhaps more importantly, a better understanding of how people resolve trust and deception issues in a wired world, and how persuasion is effected online. These issues are not well understood at this time, in large measure due to the speed with which the internet, and particularly the web, has arrived on the scene, and started to fundamentally impact human interaction.

It is for that reason that Tim Berners-Lee and others' suggestion that a science of the web be created to seek to address the impact of technology and particularly the web is so timely and important. We need to engage in social analysis of the Web. We must, in particular, learn to understand how people respond to and deal with the negative aspects of web-based social interaction so that we can seek to ameliorate them.

Note

The primary author of this paper is a legal and securities industry professional interested in multi-disciplinary aspects of web science. The author invites comments or enquiries on the topic of this paper.

NEW BRUNSWICK SECURITIES COMMISSION
COMMISSION DES VALEURS MOBILIÈRES DU NOUVEAU-BRUNSWICK

UNB

State of the Art in Trust and Reputation System: The Comparison Framework

Zeinab Noorian
University of New Brunswick

ABSTRACT

We introduce a framework for classifying and comparing trust and reputation (T&R systems). The framework dimensions encompass both hard and soft features of such systems including different witness location approaches, various reputation calculation engines, variety of information sources and rating systems, which are categorised as hard features, and also basic reputation measurement parameters, context diversity checking, reliability and honesty assessment and adaptability which are referred to as soft features. Specifically, the framework dimensions answer questions related to major characteristics of T&R systems including those parameters from the real world that should be imitated in a virtual environment.

CONTACT

Email : z.noorian@unb.ca

Adaptive Risk Management Laboratory

Director: Mihaela Ulieru

March 2010

Poster Design & Printing by Genographics® - 800.790.4001

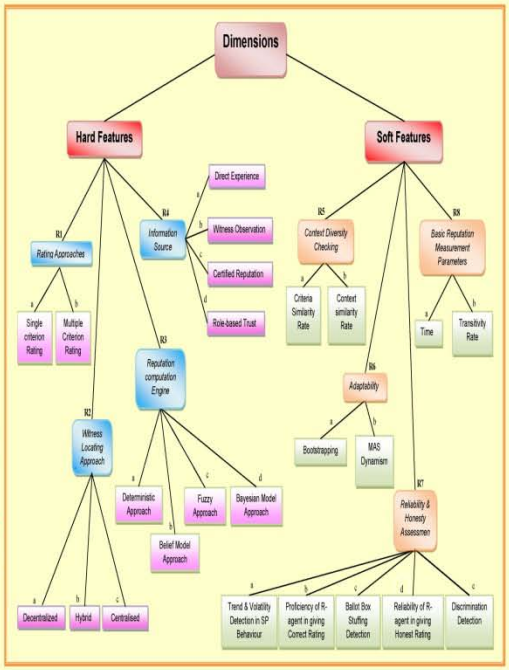
INTRODUCTION

Overcoming the inherent uncertainties and risks of the open electronic marketplace and online collaboration systems requires the establishment of mutual trust between service providers and service consumers. In fact, one of the main concerns of such environments is how the systems' resistance against self-interested participants can be enhanced and in what way their actual deceitful intentions can be understood and revealed. To address these concerns, Trust and Reputation (T&R) systems are developed to evaluate the reliability and credibility of the participants such that recommendation can be made when needed. Generally stated, the underlying goal of all T&R systems is to predict the trustworthiness and proficiency of peers in future actions based on the information gathered from their past behaviour in the environment and their peers' view towards their history. Trust can be deduced from both *individual* and *social* perspectives. Individual trust is due to direct experiences of transaction partners while social trust is calculated from third-parties experiences, which might include both honest and misleading opinions. T&R systems provide individuals with tools and techniques to deliberately solicit reputation information from peers in order to construct reasonable models of reputation for each participant. In this paper, we first give an extensive overview of five well-known trust and reputation systems. Then , we present the comparison framework and its respective dimensions. Afterwards, we thoroughly compare the existing T&R systems based on the proposed framework and analyze their pros and cons by effectively addressing some advanced features of the framework.

Summery of the Current T&R Systems

T&R System's Name	References	Distinguishing Features
FIRE	(T.D Huynh, N.R. Jennings, N.R.Shadbolt,2006)	Designed for Multi-agent system, exploits four information sources, handles the bootstrapping problem of newcomers, filters out inaccurate reputation information, attempts to differentiate between dishonest and mistaken agents, provides compound reliability measures, employs a multi-criterion rating system ,supports dynamism in open MAS.
REGRET	(Jordi Sabater and Carles Sierra, 2002)	Designed for complex e-commerce systems, develops sociogram to model social relationships, supports neighbourhood & system reputation, and provides ontological dimensions to combine various behavioural aspects of reputation. Evaluates witness honesty through fuzzy rules. Provides reliability measure; employs a multi-criteria rating system.
Model by Yu& Singh	(B. Yu, M.P. Singh,2003)	Suitable for MAS, proposes novel trust & referral network, detects three models of deceptions. Provides credibility measures pertaining to each model. Differentiate between agents having bad reputation or no reputation using Dempster-Shafter theory of evidence. Supports dynamism in open MAS.
TRAVOS	(W. T. L. Teacy, J. Patel, et al.2006)	Designed for large-scale open system, provides two information sources, exploits a probabilistic approach to determine credibility of witnesses, provides confidence metric and reliability measure for direct interaction information sources; Employs a single-rating system
PeerTrust	(L. Xiong and L. Liu,2004)	Designed for P2P e-commerce systems, provides two methods as credibility measures, supports transaction context and community context factors in trust metric, and employs an adaptive architecture for peer location. Supports dynamism in peer2peer systems. Attempts to address bootstrapping problem. Support a single-rating system.

The comparison Framework and Its Dimensions



Comparing the Current T&R Systems Using the Framework

The meanings of symbols used in the comparison Table are as follows:
N/S the model does not satisfy the corresponding feature.
P: the model attempts to address corresponding feature and has partly succeed.
Y: the model satisfied the corresponding feature.
A: the model assumes the particular feature exist and do not provide any method to address it.
N/A: the corresponding requirement is not applicable

	R1	R2	R3	R4					R5		R6		R7				R8	
	a	*b*	*a*	*b*	*c*	*a*	*b*	*c*	*d*	*a*	*b*	*a*	*b*	*a*	*b*	*c*	*d*	*e*	*a*	*b*
FIRE	Y	Y	Y			Y	Y	Y	Y			Y	Y	P	P	Y	Y	P	Y	P
REGRET	Y	A		Y		Y	Y		Y	Y		Y				Y	Y			
Model by Yu & Singh		Y		Y		Y	Y					P	Y			P	Y		Y	
TRAVOS	Y	A				Y	Y	Y		N/A	N/A					Y	Y	P	Y	
PeerTrust	Y	Y	Y			Y	Y			N/A	N/A	Y	Y	P		Y	Y	P	Y	

Conclusion

In this paper, we have introduced a framework for classifying and comparing Trust and Reputation systems and provided an overview of some prominent trust and reputation systems according to this framework pointing to ways to choose one over another for particular applications. The dimensions of this framework help system-developers to choose or build their desired T&R system with appropriate features according to their requirements. Inspired by the framework's dimensions, we intend to develop a novel trust model which can possibly satisfy the requirement of evolving environments. More explicitly, we want to introduce a decentralised adaptive model with an optimistic approach which minimise the exclusion of participants by providing suitable mechanism for differentiating between incompetence, mislead, victims of discrimination and dishonest participants.

An Online Adaptive Approach to Alert Correlation

Hanli Ren, Natalia Stakhanova, Ali A. Ghorbani

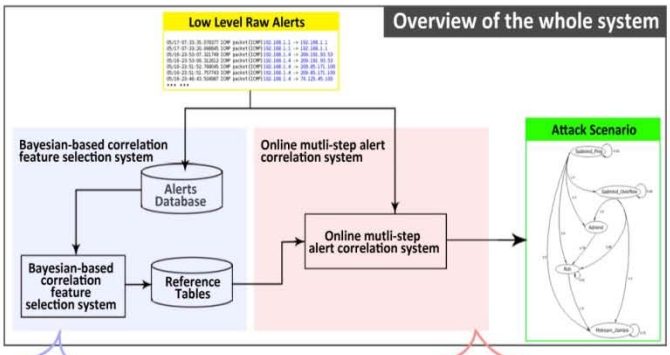
Information Security Centre of Excellence

MOTIVATION

- IDSs usually generate a tremendous number of intrusion alerts
- Alert correlation techniques aiming to provide a succinct and high-level view of attacks gained a lot of interest.
- Majority of them address the alert correlation in the off-line setting

In this work, we focus on the online approach to alert correlation. Specifically, we propose a fully automated approach for online alert correlatio.

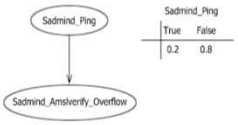
FRAMEWORK



Bayesian network

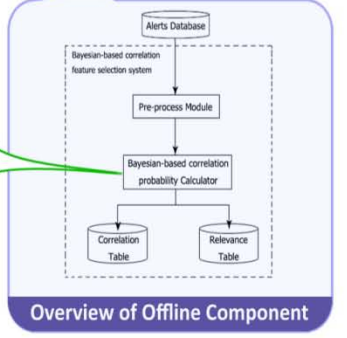
- Describe causal or dependent relationships among variables
- Illustrate the strengths of these relationships

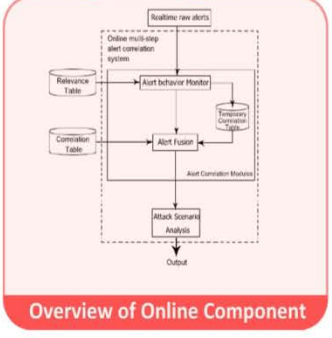
$$P(child = c|parent = p) = \frac{P(child = c \wedge parent = p)}{P(parent = p)}$$



Sadmind_Ping	True	False
	0.2	0.8

Sadmind_Ping	Sadmind_Amslverify_Overflow: True	False
True	0.9	0.1
False	0.01	0.99

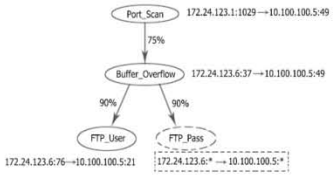




RESULTS

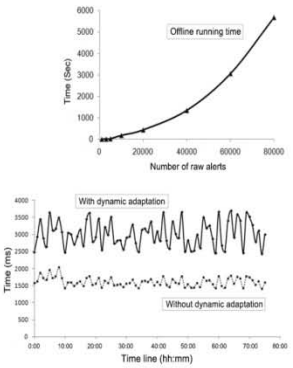
Attack scenario analysis and prediction

Alert Type Pair	Correlation Probability	Selected Features
<Port_Scan, Buffer_Overflow>	75%	DesIP,DesPort
<Buffer_Overflow, FTP_User>	90%	SrcIP,DesIP
<Buffer_Overflow, FTP_Pass>	90%	SrcIP,DesIP



An attack scenario is generated based on the pairs of causally related alerts.

Performance Test



CONTRIBUTIONS

The contributions of this work can be summarized as follows:

- ***A Bayesian correlation feature selection model*** that allows to automatically retrieve causal relationships and relevant features among alerts without expert or domain knowledge.
- ***An adaptive method for online attack scenario construction*** that allows a user to extract attack patterns in real time.
- ***An implementation*** of the proposed approach that allows a user to generate attack scenarios from a

Faculty of Computer Science
University of New Brunswick
UNB

IDS Alert Visualization for Network Security Monitoring and Analysis

alert visualization assembly

Hadi Shiravi, Ali Shiravi, Ali A. Ghorbani

Information Security Centre of Excellence

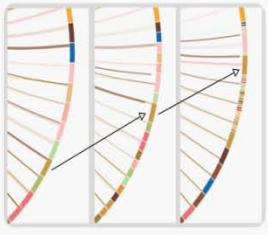
Faculty of Computer Science

www.iscx.ca

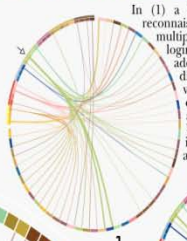
Avisa is a security visualization system that addresses the aforementioned problems. It is built upon an emerging information visualization paradigm, namely radial visualization. The paradigm is aesthetically pleasing, allows for data to be encoded on both the outer and interior parts of the ring and has a compact layout for an effortless user interaction [9].

Avisa is composed of two main components. The radial panel and the interior arcs. The radial panel itself is composed of two inner and outer rings. Starting from the top left corner, a color band inside the inner ring is used to display IDS alert types. The outer ring located exactly above this color band is used for categorizing alert types and facilitating user interaction. One color is assigned to each alert type category and different shades of the same color are used for individual alert types inside a category. We believe that this color coding eases visual correlation. The greater portion of the radial panel is devoted to internal hosts residing inside a network. Hosts can be arranged in subnets or asset groups or even be manually arranged based on specific machines that an admin is interested in monitoring. The outer ring surrounding the individual hosts (subnet panel) depicts these arrangements.

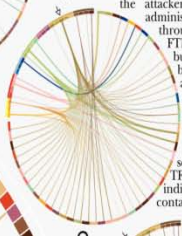
Avisa also supports filtering through direct interaction with the user. By simply clicking on any of the hosts, subnets, alert types or alert categories the entire portion of the host or alert panels are devoted to them. This feature allows for filtering of hosts and alerts in any combination.



The greatest advantage of Avisa is its support of animation. Animation is essentially a sequence of images used to convey the illusion of movement. It can facilitate perception of change over time. In our case, animation is used not only to display transitions of one view to another, but to assist in enlightening system transitions from one state to another. In Avisa we support two methods of playback. Real-time playback and delayed playback.

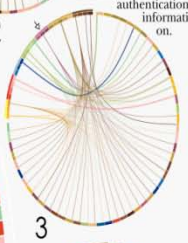


In (1) a large number of Nmap reconnaissance events have been triggered, targeting multiple hosts on the network. In (2) multiple failed login attempts are seen targeting the same range of IP addresses. Although these events are triggered at different time stamps, Avisa's time window facilitates visual correlation by displaying past and current events at the same time. The unusual number of attempts indicates that the intruder is attempting a brute force method of attack. The same procedure is seen in Figure (3) that after a brute force attempt, which is likely to have been unsuccessful, the attacker has tried to gain administrator privileges through exploiting several FTP severs vulnerable to buffer overflow. The bftpd chown overflow alert is generated on multiple servers indicating a possible intrusion. In Figure (4), the compromised FTP servers are receiving TFTP GET passwd alerts indicating a file transfer containing root level authentication information.

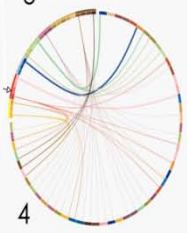


1

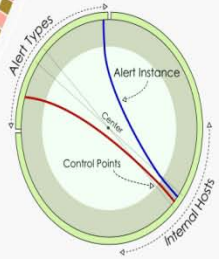
2

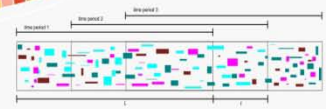


3



4





The number of alerts displayed for a particular host on the host panel is constrained by a time window whose length is specified by the user. The time window is moved every user specified period of time. In this figure on left, L specifies the windows length and t specifies the update period.

Simulating Network Intrusion in NS-2

Palash Verma, John DeDourek, Przemyslaw Pochec
Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada E3B 5A3

Introduction

▪ An Intrusion can be defined as any set of actions to compromise the integrity, confidentiality and/or availability of a resource. Intrusion attacks on Computer Networks have become a very common scenario.

▪ Denial of Service (DoS) & Distributed Denial of Service (DDoS) Intrusion attacks are very easily generated and are tough to detect as they are similar to the normal traffic packets.

▪ Network Simulation provides as a very convenient tool to model and study such intrusion attacks.

▪ In our research, we tried to observe as to how the network properties change during a DoS and DDoS attack, on a wired network.

Methodology

▪ We started by creating a model for the normal network. The network consisted of 64 wired nodes, highly meshed, to simulate a switched network, where every node is connected to the other nodes in the network. All links have 1 Mbps bandwidth, with a delay of 100 ms. The legitimate nodes are defined as UDP agents sending packets of size 600 bytes. All nodes use Constant Bit Rate (CBR) Traffic. TCL was used to model the network. A dormant attacker node is also created with 10Mbps bandwidth and 100ms delay (Node 64 visible in red hexagon on the right)

▪ Next we modeled our malicious node which is used to simulate Intrusion. This was modeled by using ICMP agent so that we can mimic a DoS attack. This node when activated sends ping messages to the compromised node, in a flood mode, i.e. as fast as possible. This is shown by red packets being sent to Node 0 on the right. Due to overwhelming ping traffic the node starts dropping packets. As a result, legitimate nodes are denied access to the compromised node.

▪ To further extend our study, we created 5 more malicious nodes (namely n65 – n69) having exactly the same characteristics as of the earlier malicious node. In this simulation, all the malicious nodes are activated together to create a DDoS attack. Once activated all the nodes send malicious ping packets in flood mode to the compromised node, denying access to other legitimate nodes.

▪ Once the modelling is completed we will run the three simulations and create respective nam files. Then we use an awk script to extract data out of the nam files

Goals

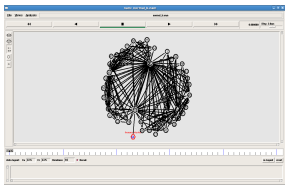
▪ We want to analyze the use of NS-2 simulations as a possible tool for studying Intrusion in networks.

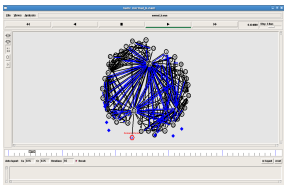
▪ We would like to consider questions such as: can we run these simulations in a reasonable time and space?

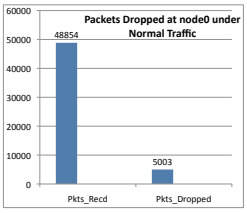
▪ Analyze the network behaviour from simulation results. Network behaviour is expected to change once an attack begins.

Experiments and Results

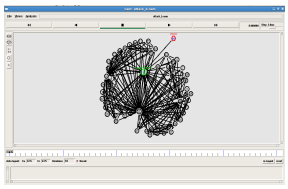
Simulation with normal traffic

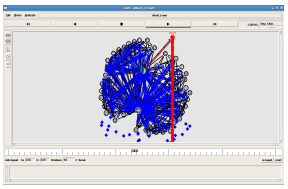


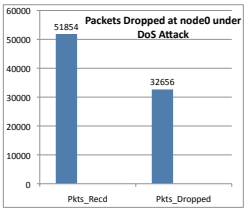




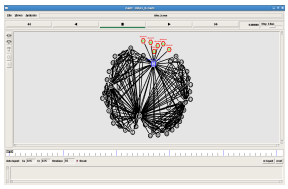
Simulation of DoS attack

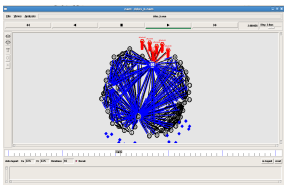


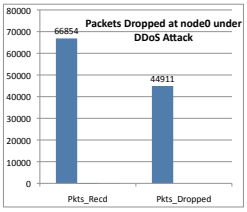




Simulation of DDoS Attack







Network Simulator *v2* (NS-2)

▪ NS-2 is a object oriented, discreet event driven network simulator, developed by UC Berkeley, written in C++ and OTcl (Tcl script language with Object-oriented Extensions). It can simulate real network structures and characteristics in the network structure.

▪ C++ defines the internals of the simulator objects where as, the OTcl is responsible for assembling, configuring and running the discreet events in the simulator environment

▪ It simulates actual network protocols such as TCP and UPD, traffic source behavior such as FTP, Telnet, Web, CBR and VBR, router queue management mechanism such as Drop Tail, RED and CBQ, routing algorithms such as Dijkstra, and much more.

▪ NS-2 also implements multicasting and some of the MAC layer protocols for LAN simulations.

▪ Once simulations are complete, NS-2 outputs either text based or animation based results.

Denial of Service (DoS)

▪ DoS is a complex and fascinating form of computer attack that impacts the confidentiality, integrity, and availability of millions of computers worldwide.

▪ The sole purpose of launching such an attack is to stop the victim computer from serving legitimate requests.

▪ Most DoS attacks exploit flaws related to the implementation of a TCP/IP model protocol.

▪ DoS can be classified in two forms namely
  • Denial of service by saturation, which involves flooding a machine with requests so it can no longer respond to actual requests
  • Denial of service by vulnerability exploitation, which involve exploiting a flaw in the remote system so as to make it unusable.

Distributed Denial of Service (DDoS)

▪ DDoS is an extensive form of DoS. In this attack multiple compromised hosts in the network attack the victim.

▪ To amplify the effect and hide real attackers, DDoS attacks can be generated in two different ways:
  • In one, the attacker compromises a number of agents and manipulates the agents to send attack traffic to the victim.
  • In other form the attacker uses reflectors. A reflector is any host that responds to a packet if it receives a packet.

▪ ICMP flooding based attack uses ICMP protocol. Usually ICMP REQUEST and ECHO REPLY messages are used for carrying control information for network management.

▪ In a typical attack the source address field of a ICMP ECHO REQUEST message is set as the victim address. Therefore, the ICMP ECHO REPLY message will be sent to the victim instead of the real request message sender (the attack agent).

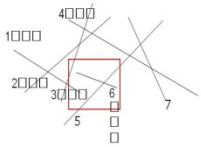
I/O-efficient Rectangular Segment Search

Gautam K. Das and Bradford G. Nickerson

Faculty of Computer science, University of New Brunswick, Fredericton, New Brunswick, Canada

Problem Definition

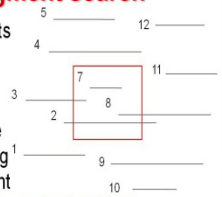
Construct a data structure to store given a set S of line segments in 2D such that an axis aligned rectangular segment query Q can be performed efficiently



Output: 1, 3, 5, 6

Restricted Segment search

-All the line segments in **S** are horizontal



-Step 1: report all the line segments having at least one end point inside query rectangle (output: 3, 7, 8)

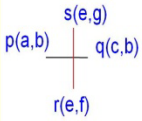
Complexities:[1]

Space - disk blocks

Query I/Os -

Step 2: Report all the segments which intersect at least one boundary of query rectangle

A (1,1,2) search (or Q(3,1) search)



The line segment [p, q] → X = (a,c,b).

The segment query is a 3D (1,1,2) point search.

Complexities of (1,1,2) search same as Step 1 [2]

Restricted case overalls complexities

Space - disk blocks, Query I/Os-

General Segment search

Step 1:Same as restricted case

Step 2: Vertical segment query

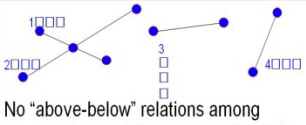
Persistent B-tree

Space - **O(N/B)** disk blocks (**N**= # updates)

Update time - **$O(\log_B N)$**

Query I/Os - **$O(\log_B N + K/B)$**

Elements are totally ordered

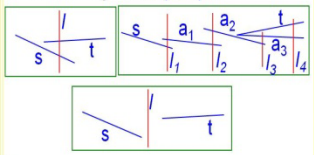


No "above-below" relations among -segments 1 and 2, segments 3 and 4

χ= Segments set induced by the end points of S and intersection points among them

$|\chi| = O(N + \lambda)$ where λ = # intersections among the segments in S

Definition: A segment s is said to be below the segment t ($s \leq t$) if



Plane sweep from left to right and

when end point encountered

If end point is start point of the segment - insert the segment in the B-tree,

else delete the segment from the B-tree

Complexities:

Persistent B-tree:

Space - $O((N+\lambda)/B)$ disk blocks

(As # of update = $O(N+\lambda)$)

Query I/Os - **$O(\log_B N + K/B)$**

Overall for general segment search:

Space - disk blocks

Query I/Os - **$O(\log_B N + K/B)$**

Entirely Within Rectangular Segment Search

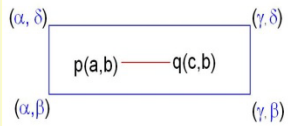
S = set of horizontal line segments in 2D

Query object = axis aligned rectangle

each segment intersecting the query is entirely within the search rectangle

Output: h

Entirely within rectangular search is a 3D (1,1,2) point search



Horizontal line (p,q) → (a,c,b)

Rectangle ABCD → Region R =

$\{(x,y,z) | \alpha \leq x, y \leq \gamma, \beta \leq z \leq \delta\}$

Complexities of (1,1,2) search [2]

Space - disk blocks

Query I/Os -

Complexities of entirely within rectangular search

Space - disk blocks

Query I/Os -

Conclusion

Open problems:

Reduce the complexity of the above problems

Consider the problems in higher dimensions

Triangular segment search problem

References: [1] L. Arge, "External memory data structures", Handbook of Massive Data Sets, J. Abello, P. M. Pardalos, M. C. G. Resende(eds.), pp. 313-358, Kluwer Academic Publishers, Dordrecht, 2002.

[2] P. Afshani, L. Arge and K.D. Larsen, "Orthogonal range reporting in three and higher dimensions, FOCS, pp. 24-27, 2009.

Sponsored by: Faculty of Computer Science UNB, NSERC CRSNG

A Data Structure for Efficient Search of Objects Moving on a Graph

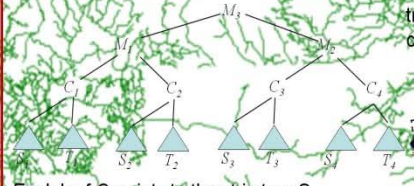
Thuy T. T. Le and Bradford G. Nickerson

Faculty of Computer science, University of New Brunswick, Fredericton, New Brunswick, Canada

Motivation

- Graph G with n moving object instances on E edges. λ intersections among moving objects paths.
- Assume objects traveling at a constant velocity over an edge.
- How to efficiently store the historical positions of moving objects to reduce the time to search for moving objects intersecting query rectangles?

Data Structure



Each leaf C_i points to the strip tree S_i representing edge e_i, and an ordered polyline tree T_i indexing moving object instances on edge e_i

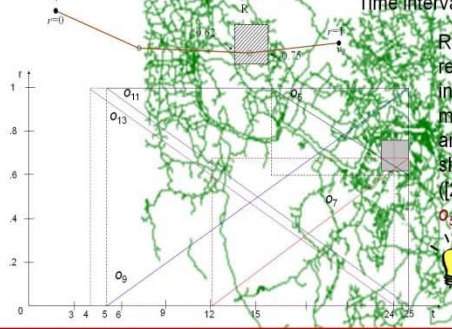
4 edges e_1,... e_4 as 4 strip trees whose bounding boxes are C_1, ..,C_4. Strip trees are merged bottom up in pairs to construct the top part of the tree.

- Balanced search tree T_i indexes line segments representing objects moving in the same direction on e_i.
- A node contains a list of points forming an ordered polyline p_i.

14 objects on e_1

id	$[t_1,t_2]$	$[r_1,r_2]$
o_1	5 10	0.5 0.8
o_2	0 9	0.7 1
o_3	10 15	0.6 0.4
o_4	5 13	0.3 0
o_5	16 25	1 0.6
o_6	2 22	0 1
o_7	12 25	0 0.68
o_8	3 23	0 1
o_9	5 25	0 1
o_{10}	1 21	0 1
o_{11}	5 25	1 0
o_{12}	0 20	1 0
o_{13}	4 24	1 0
o_{14}	3 23	1 0

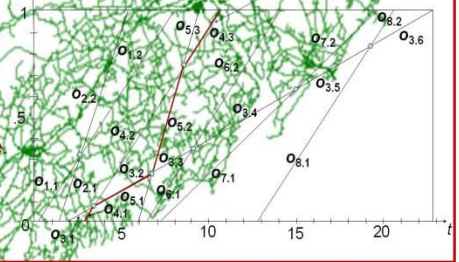
Rectangle query R intersects edge e_1 at position interval $[r_1,r_2]$=[0.62, 0.75]. Time interval query is $[t_1, t_2]$=[23, 25].



Rectangles $[t_i^1, t_i^2]$, $[r_i^1, r_i^2]$ represent moving object instances. Rectangles of moving objects o_5, o_7, o_9, o_{11}, and o_{13} intersect with the shaded query Q_2 = ([23,25],[0.62,0.75]), but only o_5 and o_7 are actually in range.

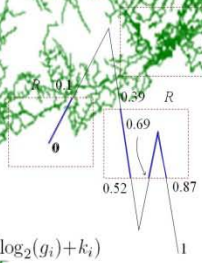
Index the diagonal line segment of rectangles

Line segments represent 8 moving objects

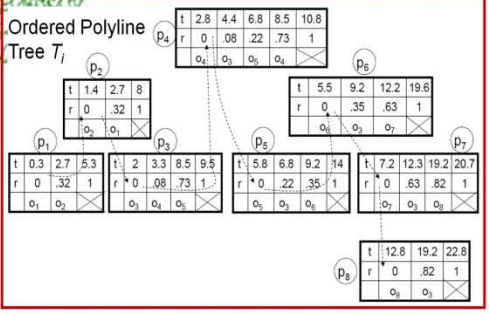


Searching

- Query Q_2 = (R, $[t_1,t_2]$), R is a rectangle, $[t_1, t_2]$ is a time interval. Q_1 is Q_2 with $t_1 = t_2$. Find moving objects intersecting Q_2.
- Start from the root to leaf nodes of the graph strip tree. R is used first to find L edges intersecting R.
- For each intersected edge e_i, ordered polyline tree T_i is used to find moving objects satisfying Q_2
- Search time for Q_1 and Q_2 queries $O(log_2 E + |L| \cdot log_2(n/E) + k)$
- Search time required for a single edge with g_i moving objects $O(\log_2(g_i) + k_i)$
- Worst case space required for the tree $O(n + \lambda + E)$, λ = O(n^2) (Pointer machine model)



Ordered Polyline Tree T_i

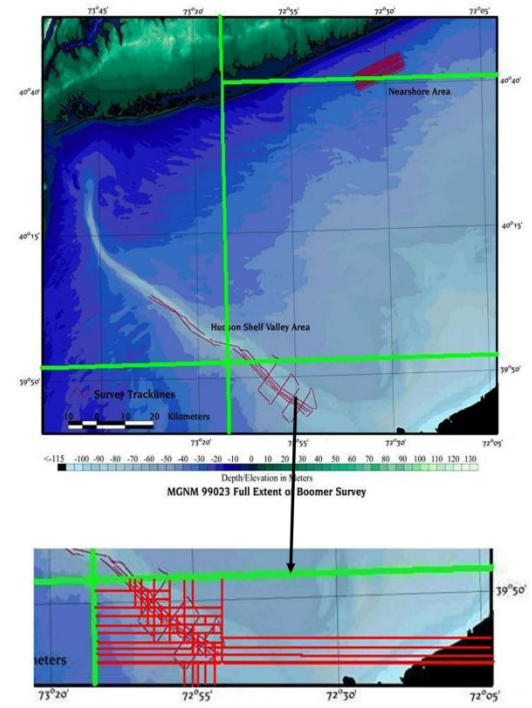


Sponsored by: Computer Science UNB, NSERC CRSNG

I/O-Efficient Spatial Data Structures: Observations on the d-Dimensional Grid File

Stuart A. MacGillivray and Bradford G. Nickerson

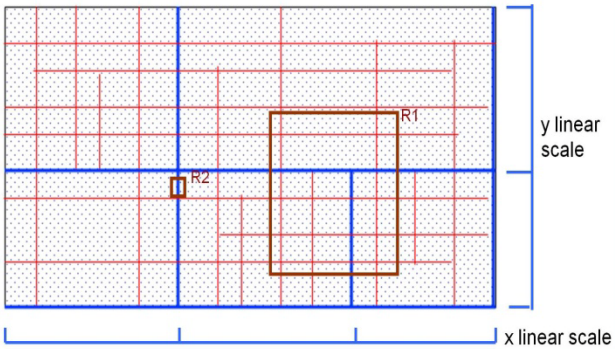
Faculty of Computer science, University of New Brunswick, Fredericton, New Brunswick, Canada



Test data and map courtesy of USGS Navigation Boomer Survey, http://quashnet.er.usgs.gov/data/1999/99023/navigation/boomer/index.htm

Structure Definition and Motivations

- The Grid File is a linear-space structure for storing multidimensional point data on a disk, allowing I/O-efficient search.
- Points are stored on the disk in cell-blocks of fixed size.
- Subdirectories stored on the disk in fixed blocks contain pointers to these cell-blocks, and linear scales describing their extent.
- Main memory M contains pointers to subdirectories and coarser linear scales.
- Retrieval of a given point is thus possible in **two disc accesses**: one to retrieve the appropriate subdirectory, and one to retrieve the block of points.
- Cells and subdirectories are spatially determined. Partitioning takes place dynamically as points are added.
- Primary motivation: Storage of large amounts of data (e.g. millions to billions of data points) and retrieval of data with minimal I/O operations
- Assumptions for tests: Disk page size of 4 kB, two-dimensional 32-bit indexing, 24 bytes per point. B = 900 blocks per subdirectory, C=170 points per block.



Range Queries and Limitations

- Range searches require few disk accesses; worst case scenario, i.e. R2, would require 2^d times as many disk accesses as necessary for retrieval of a single point.
- Range search is possible in $O(2^d + K/B)$ I/Os.
- Main directory is limited by constraints of main memory.
- Assuming main memory M of 4 GB and test assumptions, a grid file could index a maximum of 1.53×10^{14} points, around 150 terabytes of data.
- Other limitations include the structure itself; as currently written, divisions between blocks and subdirectories are determined as points are added to the file.
- Poorly distributed points may result in an inefficient structure.

Theoretical Extensions

- Extensible with additional layers of subdirectories, i.e. k layers.
- Every additional layer increases the number of disk accesses needed to retrieve a point by 1, and increases the capacity by a factor of B
- k layers store up to $N = B^k CM$ points, e.g. k=4 => $N=1.2 \times 10^{20}$
- Sufficient extension in this regard gives logarithmic time for point and range queries
- Single point retrieval: 1+k disk accesses, k at least $\log_B(\frac{N}{CM})$

References:

Jürg Nievergelt, Hans Hinterberger, and Kenneth C. Sevcik. The grid file: An adaptable, symmetric multikey file structure. *ACM Trans. Database Syst.*, 9(1):38-71, 1984

Klaus Hinrichs. Implementation of the grid file: Design concepts and experience. *BIT*, 25(4):569-592, 1985.

Sponsored by: Faculty of Computer Science UNB; NSERC CRSNG

2009 Research Publications

Title: *Customizable Bit-width in an OpenMP-based Circuit Design Tool*
Authors: T. F. Beatty, E. E. Aubanel and K. B. Kent,
Publication: 17th ACM International Symposium on Field Programmable Gate Arrays (FPGA) 2009, Monterey, USA, pp. 278, February 22-24, 2009.
Abstract

As transistor density grows, increasingly complex hard-ware designs are implemented. In order to manage this complexity, hardware design can be performed at a higher level of abstraction. High level synthesis enables the automatic conversion of algorithms into hardware imple-mentations, abstracting away the underlying complexities of hardware from the designer. A number of high level synthesis tools have recently been developed, including an OpenMP to Handel-C translator. Improvements to the translator, including a new compiler directive allowing customizable register width, are described. Using a set of benchmark tests, the OpenMP to Handel-C translator is evaluated on several criteria, with the goal of evaluating the variable bit-width effects and identifying further areas for improvement.

Title: An *OpenMP-based Circuit Design Tool: Customizable Bit-width*
Authors: T. F. Beatty, E. E. Aubanel and K. B. Kent
Piblication: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) 2009, Victoria, Canada, pp. 17-22, August 23-26, 2009.
Abstract

As transistor density grows, increasingly complex hardware designs may be implemented. In order to manage this complexity, hardware design can be performed at a higher level of abstraction. High level synthesis enables the automatic conversion of algorithms into hardware implementations, abstracting away the underlying complexities of hardware from the designer. A number of high level synthesis tools have recently been developed, including an OpenMP to HandelC translator. Improvements to the translator, including a new compiler directive allowing customizable register width, are described. Using a set of benchmark tests, the OpenMP to HandelC translator is evaluated on several criteria, with the goal of evaluating the variable bit-width effects and identifying further areas for improvement.

Title: Symmetric matroid polytopes and their generation
Authors: David Bremner, J¨urgen Bokowski, and Gabor Gevay
Publication: European Journal of Combinatorics, 30:1758–1777, November 2009.
Abstract

Matroid polytopes form an intermediate structure useful in searching for realizable convex spheres. In this article we present a class of self-polar 3-spheres that motivated research in the inductive generation of matroid polytopes, along with two new methods of generation.

Title: *Toward a unified framework for mobile applications*
Authors: Sangwhan Cha, Bernd J. Kurz, Weichang Du
Publication: Proceeding of the 9th Communication Networks and Services Research Conference (CNSR 2009), Moncton, May, 2009.
Abstract

Mobile application developers and content providers usually need to develop mobile applications with concerns for mobility for specific wireless networks and device platforms which are used by network carriers. In order to provide standard mobile applications with interoperability and mobility support, in this paper we propose a comprehensive mobile application framework to support interoperability and mobility of mobile application development and operation. Such framework supports developing mobile device applications, mobile server applications, as well as mobile client-server communications and peer-to-peer communications.

Title: *Polyhedral representation conversion up to symmetries*
Authors: David Bremner, Mathieu Dutour Sikiri´c, and Achill Sch¨urmann
Publication: Polyhedral Computation, CRM Proceedings and Lecture Notes, 48:45–71, 2009.
Abstract

We give a short survey on computational techniques which can be used to solve the representation conversion problem for polyhedra up to symmetries. We in particular discuss decomposition methods, which reduce the problem to a number of lower dimensional subproblems. These methods have been successfully used by di_erent authors in special contexts. Moreover, we sketch an incremental method, which is a generalization of Fourier-Motzkin elimination, and we give some ideas how symmetry can be exploited using pivots.

Title: *Rapid Prototyping Projection Algorithms with FPGA Technology*
Authors: J. Cole, L. E. Garey and K. B. Kent,
Pubklication: 2009 IEEE Rapid Systems Prototyping Symposium, Paris, France, pp. 95-101, June 23-26, 2009.
Abstract

Linear systems with Toeplitz coefficient matrices often appear in applied science problems. Systems of this form arise as a result of finite difference methods when applied to approximate differential Equations with boundary conditions. The sparse structure of Toeplitz matrices lend themselves well to iterative algorithms, such as projection methods, and are favored techniques for solving large systems. Field Programmable Gate Arrays (FPGAs) have been growing in popularity among the scientific community due to the potential for increased performance when evaluating mathematical operations. The regular, sparse, structure inherent in Toeplitz systems makes it suitable for FPGA acceleration. Here, a framework is developed to support the efficient development of projection algorithms in an FPGA. Results of applying the framework to two projection algorithms are presented.

Title: *Fixed-Parameter Tractability of Anonymizing Data by Suppressing Entries*
Authors: Patricia A. Evans, H. Todd Wareham, and Rhonda Chaytor
Publication: Journal of Combinatorial Optimization, Springer, 18(4):362-375 (2009).
Abstract

A popular model for protecting privacy when person-specific data is released is k-anonymity. A dataset is k -anonymous if each record is identical to at least (k−1) other records in the dataset. The basic k-anonymization problem, which minimizes the number of dataset entries that must be suppressed to achieve k-anonymity, is NP-hard and hence not solvable both quickly and optimally in general. We apply parameterized complexity analysis to explore algorithmic options for restricted versions of this problem that occur in practice. We present the first fixed-parameter algorithms for this problem and identify key techniques that can be applied to this and other k-anonymization problems.

Title: *A Design Flow for Optimal Circuit Design using Resource and Timing Estimation*
Authors: F. Gharibian and K. B. Kent,
Publication: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM) 2009, Victoria, Canada, pp. 227-233, August 23-26, 2009.
Abstract

In this paper, we study and investigate resource estimation methods that are used in circuit design for Field Programmable Gate Arrays (FPGAs). These methods usually estimate the amount of resources to be consumed by a hardware design before circuit synthesis takes place. The purpose of this study is to analyze the suitability of an estimation method for a design flow. A framework is also proposed to help the optimization process of the design. This framework automatically optimizes the design by finding potential parallelism in the design and applies it while considering the available resource and time constraints.

Title: *A Hardware/Software Co-specification Methodology Based Upon OpenMP*
Authors: T. S. Hall and K. B. Kent
Publication: IEEE Toronto International Conference – Science and Technology for Humanity, Symposium on Electronic Design Automation 2009, Toronto, Canada, pp. 714-719, September 27 - 29, 2009.
Abstract

This paper presents a hardware/software co-specification methodology based on the OpenMP parallel programming specification. The methodology sets out the procedures to convert a system specified as an OpenMP software application into a hardware/software design. The methodology is intended to permit software developers to produce custom hardware/software system specifications using software development tools.

Title: *To oli Gate Implementation Using The Billiard Ball Model*
Authors: H. Hosseini, G. Dueck
Publication: 40th International Symposium on Multiple-Valued Logic, May 26-28, 2010, Casa Convalescncia, Barcelona, Spain
Abstract

In this paper we review the Billiard Ball Model (BBM) introduced by Toffoli and Fredkin. The analysis of a previous approach to design reversible networks based on BBM it shown to

ignored physical realities. We prove that some logic function cannot be realized without additional control balls. For example, to realize the logical OR operation, at least three control balls are needed. We show how reversible Toffoli gates can be constructed with this model. Finally, a Toffoli gate module is proposed that can be used in a cascade of gates and thus implement arbitrary reversible functions.

Title: *Modeling and simulation of norms and institutions in multi-organizational systems using BDI framework*
Authors: H.Hosseini, M.Ulieru
Publication: APICS 2009, Dalhousie University, Halifax, NS, Canada.
Abstract

In recent years, modeling and designing more complex structures of the socio-technical systems have been a vital question to address in multi-agent system society. Solving a complex system and leading it toward a certain goal can be done using multi-agent systems, as it is becoming a promising new programming paradigm called Agent-Oriented Software Engineering (AOSE), where there are plenty of agents collaborating with each other to achieve a common goal. Coordination and communication in multi-organizational systems is considered to be complicated to handle due to the various institutional needs of rules and norms in every organization. In this paper, we would like to present a methodology to design and implement policies and norms inside each organization by a logic-based approach to show all the rules and regulations used by the agents considering autonomy and pro-activity of the agents. We will show the necessary steps to implement the conceptual level of interaction between agents in the BDI (belief-desire-Intention) framework using Brahms simulation tool. Later, we would like to discuss difficulties in managing policies on meta-organizational scale by observing the best practices of individuals and also groups in the field, and define a possible solution to overcome the decision-making process and optimizing the interactions in a crisis scenario.

Title: *An Autonomous Agent-based Framework for Self-Healing Power Grid*
Authors: H. Hosseini, Z.Noorian, M.Ulieru,
Publication: IEEE Systems, Man, and Cybernetics 2009 Conference, October 11-14, San Antonio, Texas, USA.
Abstract

Reliable, secure and robust power grid network is a necessity for crucial financial, industrial and business networks. Since national electrical grid, telecommunication, information networks and transportation networks are interdependent critical infrastructures, having an agent-based self-healing framework to reduce cascading failures through the networks and finding reasonable solution for potential faults – would be an essential asset. In response to this need we propose a self-healing framework that employs advanced failure diagnosis techniques along with autonomous web services to provide temporary recovery solutions. Furthermore, it provides a cognitive planning cycle to find ultimate corrective solutions as well as evaluation service to verify the effectiveness and performance of the final solution.

Title: *Determining the Optimal FPGA Design for Computing Highly Parallel Problems*
Authors: K. B. Kent and J. E. Rice
Publication: IET Computer and Digital Techniques journal, vol. 3, issue 3, pp. 247-258, 2009.
Abstract

Reconfigurable hardware has recently shown itself to be an appropriate solution to speeding up problems that are highly dependent on a particular complex or repetitive sub-algorithm. In most cases these types of solutions lend themselves well to parallel solutions. We investigate the optimal design, maximizing performance while existing within the target FPGA resources, on FPGAs for problems with algorithms or sub-algorithms that can be highly parallelized.

Title: *A Novel Bayes Model: Hidden Naive Bayes*
Authors: Liangxiao Jiang, Harry Zhang, Zhihua Cai
Publication: IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 10, 2009, pp. 1361-1371
Abstract

Because learning an optimal Bayesian network classifier is an NP-hard problem, learning-improved naive Bayes has attracted much attention from researchers. In this paper, we summarize the existing improved algorithms and propose a novel Bayes model: hidden naive Bayes (HNB). In HNB, a hidden parent is created for each attribute which combines the influences from all other attributes. We experimentally test HNB in terms of classification accuracy, using the 36 UCI data sets selected by Weka, and compare it to naive Bayes (NB), selective Bayesian classifiers (SBC), naive Bayes tree (NBTree), tree-augmented naive Bayes (TAN), and averaged one-dependence estimators (AODE). The experimental results show that HNB significantly outperforms NB, SBC, NBTree, TAN, and AODE. In many data mining applications, an accurate class probability estimation and ranking are also desirable. We study the class probability estimation and ranking performance, measured by conditional log likelihood (CLL) and the area under the ROC curve (AUC), respectively, of naive Bayes and its improved models, such as SBC, NBTree, TAN, and AODE, and then compare HNB to them in terms of CLL and AUC. Our experiments show that HNB also significantly outperforms all of them.

Title: *LEARNING DECISION TREES WITH LOG CONDITIONAL LIKELIHOOD*
Authors: HAN LIANG, YUHONG YAN, HARRY ZHANG
Publication: International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), Volume: 24, Issue: 1(2010) pp. 117-151
Abstract

In machine learning and data mining, traditional learning models aim for high classification accuracy. However, accurate class probability prediction is more desirable than classification accuracy in many practical applications, such as medical diagnosis. Although it is known that decision trees can be adapted to be class probability estimators in a variety of approaches, and the resulting models are uniformly called Probability Estimation Trees (PETs), the performances of these PETs in class probability estimation, have not yet been investigated. We begin our research by empirically studying PETs in terms of class probability estimation, measured by Log Conditional Likelihood (LCL). We also compare a PET called C4.4 with other representative models, including NaÃ¯ve Bayes, NaÃ¯ve Bayes Tree, Bayesian Network, KNN and SVM, in LCL. From our experiments, we draw several valuable conclusions. First, among various tree-

based models, C4.4 is the best in yielding precise class probability prediction measured by LCL. We provide an explanation for this and reveal the nature of LCL. Second, compared with non tree-based models, C4.4 also performs best. Finally, LCL does not dominate another well-established relevant metric â€" AUC, which suggests that different decision-tree learning models should be used for different objectives. Our experiments are conducted on the basis of 36 UCI sample sets. We run all the models within a machine learning platform â€" Weka. We also explore an approach to improve the class probability estimation of NaÃ¯ve Bayes Tree. We propose a greedy and recursive learning algorithm, where at each step, LCL is used as the scoring function to expand the decision tree. The algorithm uses NaÃ¯ve Bayes created at leaves to estimate class probabilities of test samples. The whole tree encodes the posterior class probability in its structure. One benefit of improving class probability estimation is that both classification accuracy and AUC can be possibly scaled up. We call the new model LCL Tree (LCLT). Our experiments on 33 UCI sample sets show that LCLT outperforms all state-of-the-art learning models, such as NaÃ¯ve Bayes Tree, significantly in accurate class probability prediction measured by LCL, as well as in classification accuracy and AUC.

Title: *An Embedded Implementation of the Common Language Infrastructure*
Authors: J. C. Libby and K. B. Kent,
Publication: Elsevier Journal of System Architectures, vol. 55, pp. 114-126, February 2009.
Abstract

The Common Language Infrastructure provides a unified instruction set which may be targeted by a variety of high level language compilers. This unified instruction set simplifies the construction of compilers and gives application designers the ability to choose the high level programming language that best suits the problem being solved. Many compilers that target the Common Language Infrastructure exists today including Microsoft's suite of compilers provided with Visual Studio .NET. While the Common Language Infrastructure solves many problems related to design of applications and compilers, it is not without its own problems. The Common Language Infrastructure is based upon a virtual machine, much like the Java Virtual Machine. This requires that all instructions being executed on the Common Language Infrastructure be translated to native machine instructions before they can be executed on the host processor. This leads to degradation in performance as time must now be spent performing this translation. In order to overcome this problem it is proposed that an embedded processor capable of natively executing the CLI instruction set be developed. Natively executing the Common Language Infrastructure instructions will remove the need for translation to a general purpose instruction set and may increase the performance of applications executing on the proposed hardware platform. Rapid adoption of languages that target the CLI, along with growth in the field of embedded systems provides justification for exploring the feasibility of creating an embedded processor capable of executing the Common Language Infrastructure instruction set. The objective of this work is the design and implementation, using VHDL and simulation, of an embedded processor capable of natively executing the CLI instruction set. This processor provides a platform easily targeted by software developers.

Title: A Methodology for Rapid Optimization of HandelC Specifications
Authors: J. C. Libby and K. B. Kent,
Publication: 2009 IEEE Rapid Systems Prototyping Symposium, Paris, France, pp. 81-87, June 23-26, 2009.

Abstract

Utilizing high level hardware description languages for the creation of customized circuits facilitates the rapid development and deployment of new hardware. While hardware design languages increase the speed at which hardware can be developed, creating hardware designs that are both efficient in resource usage and processing speed can be time consuming and require much experience. This problem is compounded more by the long design cycle times that are introduced by the long compilation and synthesis times that are required to translate a high level hardware description language to a circuit. This problem is addressed by performing some of the optimizations automatically, pre-synthesis, reducing the total number of synthesis cycles that are required, saving much development time.

Title: *Automatic Discovery of Botnet Communities on Large-Scale Communication Networks*
Authors: Wei Lu, Mahbod Tavallaee, and Ali A. Ghorbani
Pubication: Proceedings of the 2009 ACM Symposium on InformAtion, Computer and Communications Security (ASIACCS'09), pp. 1-10, March 2009.

Abstract

Botnets are networks of compromised computers infected with malicious code that can be controlled remotely under a common command and control (C&C) channel. Recognized as one the most serious security threats on current Internet infrastructure, advanced botnets are hidden not only in existing well known network applications (e.g. IRC, HTTP, or Peer-to-Peer) but also in some unknown or novel (creative) applications, which makes the botnet detection a challenging problem. Most current attempts for detecting botnets are to examine traffic content for bot signatures on selected network links or by setting up honeypots. In this paper, we propose a new hierarchical framework to automatically discover botnets on a large-scale WiFi ISP network, in which we first classify the network traffic into different application communities by using payload signatures and a novel cross-association clustering algorithm, and then on each obtained application community, we analyze the temporal-frequent characteristics of flows that lead to the differentiation of malicious channels created by bots from normal traffic generated by human beings. We evaluate our approach with about 100 million flows collected over three consecutive days on a large-scale WiFi ISP network and results show the proposed approach successfully detects two types of botnet application flows (i.e. Blackenergy HTTP bot and Kaiten IRC bot) from about 100 million flows with a high detection rate and an acceptable low false alarm rate.

Title: *BotCop: An Online Botnets Traffic Classifier*
Authors: Wei Lu, Mahbod Tavallaee, Goaletsa Rammidi and Ali A. Ghorbani
Publication: Proceedings of Seventh Annual Conference on Communication Networks and Services Research (CNSR'09), pp. 70-77, May 2009.

Abstract

A botnet is a network of compromised computers infected with malicious code that can be controlled remotely under a common command and control (C&C) channel. As one the most

serious security threats to the Internet, a botnet cannot only be implemented with existing network applications (e.g. IRC, HTTP, or Peer-to-Peer) but also can be constructed by unknown or creative applications, thus making the botnet detection a challenging problem. In this paper, we propose a new online botnet traffic classification system, called BotCop, in which the network traffic are fully classified into different application communities by using payload signatures and a novel decision tree model, and then on each obtained application community, the temporal-frequent characteristic of flows is studied and analyzed to differentiate the malicious communication traffic created by bots from normal traffic generated by human beings. We evaluate our approach with about 30 million flows collected over one day on a large-scale WiFi ISP network and results show that the proposed approach successfully detects an IRC botnet from about 30 million flows with a high detection rate and a low false alarm rate.

Title: *Hybrid Traffic Classification Approach Based on Decision Tree*
Authors: Wei Lu, Mahbod Tavallaee and Ali A. Ghorbani
Publication: Proceedings of the 2009 IEEE Global Telecommunications Conference (GLOBECOM'09), pp. 1-6, December 2009.

Abstract

Classifying network traffic is very challenging and is still an issue yet to be solved due to the increase of new applications and traffic encryption. In this paper, we propose a novel hybrid approach for the network flow classification, in which we first apply the payload signature based classifier to identify the flow applications and unknown flows are then identified by a decision tree based classifier in parallel. We evaluate our approach with over 100 million flows collected over three consecutive days on a large-scale WiFi ISP network and results show the proposed approach successfully classifies all the flows with an accuracy approaching 93%.

Title: *Examining Implementations of a Computationally Intensive Problem in GF(3)*
Authors: J. Lutes, J. C. Libby and K. B. Kent
Publication: International Journal On Advances in Software, vol 2., no. 1, issn 1942-2628, pp. 119-130, May 2009.

Abstract

Computing the irreducible and primitive polynomials under GF(3) is a computationally intensive task. A hardware implementation of this algorithm should prove to increase performance, reducing the time needed to perform the computation. Previous work explored the viability of a co-designed approach to this problem and this work continues addressing the problem by moving the entire algorithm into hardware. Handel-C was chosen as the hardware description language for this work due to its similarities with ANSI C used in the software implementation. A hardware design for the algorithm was developed and optimized using several different optimizations techniques before arriving at a final design.

Title: *Online Classification of Network Flows*
Authors: Mahbod Tavallaee, Wei Lu, and Ali A. Ghorbani,
Publication: Proceedings of Seventh Annual Conference on Communication Networks and Services Research (CNSR'09), pp. 78-85, May 2009.
Abstract

Online classification of network traffic is very challenging and still an issue to be solved due to the increase of new applications and traffic encryption. In this paper, we propose a hybrid mechanism for online classification of network traffic, in which we apply a signature-based method at the first level, and then we take advantage of a learning algorithm to classify the remaining unknown traffic using statistical features. Our evaluation with over 250 thousand flows collected over three consecutive hours on a largescale ISP network shows promising results in detecting encrypted and tunneled applications compared to other existing methods.

Title: *A Detailed Analysis of the KDD CUP 99 Data Set*
Authors: Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani
Publication: Proceedings of the 2009 IEEE Symposium Computational Intelligence for Security and Defense Applications (CISDA'09), July 2009.
Abstract

During the last decade, anomaly detection has attracted the attention of many researchers to overcome the weakness of signature-based IDSs in detecting novel attacks, and KDDCUP'99 is the mostly widely used data set for the evaluation of these systems. Having conducted a statistical analysis on this data set, we found two important issues which highly affects the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, we have proposed a new data set, NSL-KDD, which consists of selected records of the complete KDD data set and does not suffer from any of mentioned shortcomings.

Title: *Toffoli Gate Cascade Generation Using ESOP Minimization and QMDD-based Swapping*
Authors: J. E. Rice, M. A. Thornton, K. Fazel and K. B. Kent,
Publication: 2009 Reed-Muller Workshop, Okinawa, Japan, pp. 63-72, May 23-24, 2009.
Abstract

Two methods for Toffoli gate cascade synthesis of reversible logic circuits are presented. One is based on the authors previous work, utilizing an ESOP minimization technique and then applying template-matching. The other is based on a QMDD representation of a Toffoli cascade and determining an ordering that implements the desired function. Experimental results are presented showing the feasibility of both techniques.

Title: *Protocol: A Controller for First-Responder Ecosystems*
Author: William Ross
Publication: 35th Annual Conference of IEEE Industrial Electronics (Special Session on Digital Ecosystems), 2009, pp. 3948-3953
Abstract

In recent years, much literature has surfaced regarding ecosystems; however, the nature of the interaction between various members of these ecosystems and how the interaction can be improved via organizational structures have remained relatively unexplored. In this paper, a

survey of recent work related to business survival in the current Information Age is presented in the context of First-Responder Ecosystems. This information is synthesized in such a way as to contribute to the ongoing discussion of how members from different organizations can more easily collaborate and over time, eventually, self-organize. The approach suggested in this paper is based on the success of Internet protocol in maintaining decentralized control. This paper proposes a similar approach as an alternative to the classic hierarchical organizational structure. This innovative approach creates a synergistic environment in which community, rather than individualism, is stressed and rewarded. To compare this approach with existing ones, an agent-based simulation is being planned.

Title: *Mapping Transcription Factors from a Model to a non-Model Organism*
Authors: Rachita Sharma, Patricia A. Evans, Virendrakumar C. Bhavsar
Publication: Proceedings of the International Joint Conferences on Bioinformatics, Systems Biology and Intelligent Computing, 2009, pp.160-166.
Abstract:

Identification of regulatory elements, such as transcription factors, is useful in construction of regulatory networks and to understand gene regulation. These transcription factors have already been recognized for model organisms based on extensive experiments but have not been as heavily investigated for non-model organisms. This paper proposes to use Basic Local Alignment Search Tool (BLAST) to map the transcription factors from a model to a non-model organism. Experiments are performed on bacterial organisms based on evolutionary distance to compare the results. Analysis of the results suggests that transcription factors can be mapped from one bacterial organism to another as transcription factor motifs are well preserved among these organisms. Results are also analyzed to determine the best suitable threshold for the e-value parameter of BLAST that can be used to map transcription factors, determine to be the e-value thresholds of 0.01 and 0.1. Both the BLAST e-value threshold and evolutionary distance from the model organism used for mapping have significant impact on the quality of results.

Title: *Transcription Factor mapping between Bacteria Genomes*
Authors: Rachita Sharma, Patricia A. Evans, Virendrakumar C. Bhavsar
Publication: International Journal of Functional Informatics and Personalised Medicine, 2009, Vol. 2-4, pp. 424-441
Abstract:

Identification of gene regulatory networks is useful in understanding gene regulation in any organism. Some regulatory network information has already been determined experimentally for model organisms, but much less has been identified for non-model organisms, and the limited amount of gene expression data available for non-model organisms makes inference of regulatory networks difficult. This paper proposes a method to map the regulatory links from a model to a non-model organism. Mapping a regulatory network involves mapping the transcription factors and target genes from one genome to another. In the proposed method, Basic Local Alignment Search Tool (BLAST) and InterProScan are used to map the transcription factors, whereas BLAST along with the transcription factor binding site motifs are used to map the target genes. Experiments are performed to map the regulatory network data of Saccharomyces cerevisiae to Arabidopsis thaliana. Since limited information is available about gene regulatory network links, gene expression data is used to analyze results. A set of rules are

defined on the gene expression experiments to identify the predicted regulatory links that are well supported. It is shown that more than two-thirds of the predicted regulatory links that were analyzed using gene expression data have been verified as correctly mapped regulatory links in the target genome.

Title: Dynamic Parallelization for RNA Structure Comparison
Authors: Eric Snow, Eric Aubanel, and Patricia Evans
Publication: Proceedings of the Eighth IEEE International Workshop on High Performance Computational Biology, May 2009. 8 pages.

Abstract

In this paper we describe the parallelization of a dynamic programming algorithm used to find common RNA secondary structures including pseudoknots and similar structures. The sequential algorithm is recursive and uses memoization and data-driven selective allocation of the tables, in order to cope with the high space and time demands. These features, in addition to the irregular nature of the data access pattern, present particular challenges to parallelization. We present a new manager-worker approach, whereworkers are responsible for task creation and the manager's sole responsibility is overseeing load balancing. Special considerations are given to the management of distributed, dynamic task creation and data structures, along with general inter-process communication and load balancing on a heterogeneous computational platform. Experimental results show a modest level of speedup with a highly-scalable level of memory usage, allowing the comparison of much longer RNA molecules than is possible in the sequential implementation.

Title: *Multi-service load sharing for resource management in the cellular/WLAN integrated network*
Authors: W. Song and W. Zhuang,
Publication: *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 725-735, Feb. 2009.

Abstract:

With the interworking between a cellular network and wireless local area networks (WLANs), an essential aspect of resource management is taking advantage of the overlay network structure to efficiently share the multi-service traffic load between the interworked systems. In this study, we propose a new load sharing scheme for voice and elastic data services in a cellular/WLAN integrated network. Admission control and dynamic vertical handoff are applied to pool the free bandwidths of the two systems to effectively serve elastic data traffic and improve the multiplexing gain. To further combat the cell bandwidth limitation, data calls in the cell are served under an efficient service discipline, referred to as shortest remaining processing time (SRPT). The SRPT can well exploit the heavy-tailedness of data call size to improve the resource utilization. An accurate analytical model is developed to determine an appropriate size threshold so that data calls are properly distributed to the integrated cell and WLAN, taking into account the load conditions and traffic characteristics. It is observed from extensive simulation and numerical analysis that the new scheme significantly improves the overall system performance.

Title: *Performance analysis and enhancement of cooperative retransmission strategy for delay-sensitive real-time services*
Authors: W. Song and W. Zhuang
Publication: *Proceedings of IEEE Global Communications Conference (GLOBECOM'09)*, Nov. 2009.
Abstract:

As a very promising technique, multi-hop relay has been considered in many wireless networks. It can take advantage of the inherent broadcasting nature of wireless transmission and facilitate cooperative communications. In this paper, we develop an effective analytical framework to study the delay performance of cooperative retransmission strategies. All neighbour nodes overhearing the in-progress transmission cooperate in a distributed manner and contribute to retransmissions. In particular, we focus on the application of cooperative retransmission for delay-sensitive real-time services. Based on the proposed analytical framework, the cumulative distribution function of packet transfer delay can be numerically evaluated. Accordingly, we investigate the delay outage probability (i.e., the probability of violating the maximum delay bound), which is an essential statistical quality-of-service (QoS) metric for real-time services. Further, an enhancement approach is proposed to reduce unnecessary power consumption on retransmissions. It dynamically adapts the transmission probabilities of all participating nodes, depending on current retransmission count. As shown in the numerical results, the adaptive cooperative strategy can achieve a better trade-off between satisfying delay constraint and minimizing total power consumption.

Title: *Adaptive packetization for error-prone transmission over 802.11 WLANs with hidden terminals*
Authors: W. Song, M. N. Krishnan, and A. Zakhor,
Publications: *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP'09)*, Oct. 2009 (Top 10% Award).

Abstract:

Collision and fading are the two main sources of packet loss in wireless local area networks (WLANs) and as such, both are affected by the packetization at the medium access control (MAC) layer. While a larger packet is preferred to balance protocol header overhead, a shorter packet is less vulnerable to packet loss due to channel fading errors or staggered collisions in the presence of hidden terminals. Direct collisions due to backoff are not affected by packet size. Recently, Krishnan et. al. have developed a new technique for estimating probabilities of various components of packet loss, namely, direct and staggered collisions and fading. Motivated by this work, in this paper, we exploit ways in which packetization can be used to improve throughput performance of WLANs. We first show analytically that the effective throughput is a unimodal function of the packet size when considering both channel fading and staggered collisions. We then develop a measurement-based algorithm based on golden section search to arrive at an optimal packet size for MAC-layer transmissions. Our simulations demonstrate that packetization based on our search algorithm can greatly improve the effective throughput of sensing-limited nodes, and reduce video frame transfer delay in WLANs.

Title: *Performance evaluation of interactive data services under sharing and preemptive scheduling disciplines*
Authors: W. Song, W. Zhuang, and D. Zhao
Publication: *Proceedings of IEEE International Conference on Communications (ICC'09)*, June 2009.
Abstract:

As specified by the third-generation (3G) wireless networks such as the universal mobile telecommunication system (UMTS), interactive data services, such as Web browsing, voice messaging, and file transfer, represent a major service class in operation nowadays. In this paper, we develop an analytical approach to evaluate the performance of interactive data services under sharing and preemptive scheduling. Specifically, we take into account user interactions in data sessions and the heavy-tailed data file size. Both the mean and the standard deviation of data transfer delay are investigated for the two representative scheduling disciplines. Numerical results are given to show the validity of the evaluation approach and the impact of the on-off user behaviour under the scheduling disciplines.

Title: *Knowledge representation and consistency checking in a norm-parameterized fuzzy description logic*
Authors: J. Zhao, H. Boley and W. Du,
Publication: Proceedings of the International Conference on Intelligent Computing (ICIC 2009), LNAI 5755, 2009, pp. 111-123.
Abstract

This paper has its motivation in the occurrence of uncertain knowledge in different application areas, and introduces an expressive fuzzy description logic that extends classical description logics to many-valued logics. We represent, and reason with, uncertain knowledge in the description logic ALCHIN extended by an interval-based, norm-parameterized Fuzzy Logic. First, the syntax and the semantics of the proposed fuzzy description logic are addressed. Then the paper presents an algorithm for consistency checking of knowledge bases in the proposed language.

Title: *A reasoning procedure for the fuzzy description logic fALCHIN*
Authors: J. Zhao and H. Boley,
Publication: Proceedings of the Second Canadian Semantic Web Working Symposium, 2009, pp. 46-59.
Abstract

This paper introduces an expressive fuzzy description logic that extends classical description logics to many-valued logics. This proposed fuzzy description logic extends the expressiveness of the well know description logic by fuzzy concepts, fuzzy roles, fuzzy axioms, fuzzy inverse roles, and fuzzy role inclusion axioms, as well as fuzzy at-most/at least number restrictions. This paper focuses on presenting an extended tableau algorithm for reasoning with knowledge bases in the proposed fuzzy description logic.

Title: Dynamic Parallelization for RNA Structure Comparison
Authors: Eric Snow, Eric Aubanel, and Patricia Evans
Publication: Proceedings of the Eighth IEEE International Workshop on High Performance Computational Biology, May 2009. 8 pages.

Abstract

In this paper we describe the parallelization of a dynamic programming algorithm used to find common RNA secondary structures including pseudoknots and similar structures. The sequential algorithm is recursive and uses memoization and data-driven selective allocation of the tables, in order to cope with the high space and time demands. These features, in addition to the irregular nature of the data access pattern, present particular challenges to parallelization. We present a new manager-worker approach, whereworkers are responsible for task creation and the manager's sole responsibility is overseeing load balancing. Special considerations are given to the management of distributed, dynamic task creation and data structures, along with general inter-process communication and load balancing on a heterogeneous computational platform. Experimental results show a modest level of speedup with a highly-scalable level of memory usage, allowing the comparison of much longer RNA molecules than is possible in the sequential implementation.

2009 PhD Theses

Title: *Semi-supervised Learning and Opinion-oriented Information Extraction*
Student: Bin Wang
Supervisors: Dr. Huajie Zhang, Dr. Bruce Spencer
Abstract

Recently, information extraction (IE) attracts much attention in the research of natural language processing (NLP). More and more people are not satisfied only extracting the factual information so that the study of opinion- oriented IE has become a promising research area. However, the requirement for large manually labeled training corpora is widely recognized as a bottleneck in the use of machine learning algorithms for opinion-oriented IE. As a result, semi-supervised learning is regarded as a good solution to break this bottleneck. A semi-supervised learning algorithm learns a model based on a relatively small amount of labeled instances and a large amount of unlabeled instances. In this thesis, we mainly focus on three types of semi-supervised learning algorithms, namely, self-training, co-training, and graph- based methods. For self-training, we apply value difference metric (VDM) as the selection metric and use naive Bayes and decision tree algorithms as underlying classifiers. For co-training, we propose an unsymmetrical co-training algorithm which combines an EM classifier and a self-training classifier together in an unsymmetrical structure without splitting the attribute set. For graph-based methods, we put forward a probability propagation algorithm based on the instance-attribute graph, where there are two kinds of nodes, i.e., instance nodes and attribute nodes; and two types of messages, i.e., instance node messages and attribute node messages. The goal of probability propagation algorithm is to propagate messages between nodes in order to balance the global and local situations and smooth the graph. From the experimental results, the new techniques and novel algorithms achieve better performances against their corresponding opponents. Furthermore, some of opinion-oriented IE tasks have been tackled by the semi-supervised learning algorithms in this thesis. We mainly focus on three tasks, that is, sentence subjectivity classification, contextual polarity recognition, and opinion entity identification. Particularly, self-training is used to solve the sentence subjectivity classification, co-training is used to deal with the contextual polarity recognition, and graph-based methods are used to tackle the opinion entity recognition. The experiments have been designed to compare the performances of corresponding algorithms to the corresponding tasks. The results show that semi-supervised learning algorithms are suitable for the tasks of opinion-oriented IE. Especially, the proposed techniques and algorithms also outperform their opponents in these tasks.

Title: *Combining Intensification and Diversification in Local Search for SAT*
Student: Wanxia Wei
Supervisors: Dr. Chu Min Li, Dr. Huajie Zhang
Abstract

The propositional satisfiability problem (SAT) is to determine whether there is a truth assignment to Boolean variables such that every clause in conjunctive normal form is satisfied. Intensification refers to search strategies that intend to greatly improve a solution, while diversification refers to search strategies that help achieve a reasonable coverage in the search space. Roughly speaking, there are three classes of local search algorithms for solving SAT: non-weighting, clause weighting, and variable weighting. A non-weighting algorithm mainly focuses on intensifying the search. A clause weighting algorithm and a variable weighting algorithm use clause and variable weighting, respectively, to diversify the search. One way to design a local search algorithm that is effective on many types of instances is allowing it to switch among heuristics in order to combine the strengths of these heuristics and eliminate the weaknesses of them. Two novel switching criteria, namely the evenness or unevenness of distribution of variable weights and the evenness or unevenness of distribution of clause weights, are proposed. These criteria are applied to our newly proposed heuristics and state-of-the-art heuristics. The resulting local search algorithms are *FH* (Four Heuristics), *Hybrid*, and *NCVW* (Non-, Clause, and Variable Weighting), each of which switches among two to four heuristics according to one or two of the proposed criteria in order to intensify or diversify the search when necessary. *FH*, *Hybrid*, and *NCVW* have two main characteristics in common. First, they all combine intensification strategies with suitable diversification strategies. Second, they all intensify or diversify the search when necessary. Experimental results show that, on a wide range of instances, each of algorithms *FH*, *Hybrid*, and *NCVW* exhibits generally better performance than its constituent heuristics. In addition, on a wide range of instances, each algorithm is generally effective, while state-of-the-art local search algorithms that include the best local search algorithms in the international SAT 2005 or SAT 2007 competitions, are not. Moreover, on representative instances, each algorithm shows better overall performance than *10* state-of-the-art local search algorithms, including the best local search algorithms in the international SAT 2005 and SAT 2007 competitions.

2009 Master of Computer Science Theses and Reports

Title: *Generating Partial COP-nets on Demand*
Student: Henry Bediako-Asare
Supervisors: Dr. Michael Fleming, Dr. Scott Buffett
Abstract

The Conditional Outcome Preference Network, also known as a COP-net, has been developed to graphically represent the model of a user's preferences over a set of possible outcomes. Typically, the number of elicited preferences upon which to construct a COPnet is limited. The structure of these partial preferences is then used to predict preferences over an entire set of possible outcomes. The existing methodology for constructing a COP-net includes all possible outcomes and grows exponentially in the number of attributes that describe the outcomes, thus making the construction of the COP-net infeasible. In this thesis, a different approach for constructing COP-nets, using A* search, is introduced. With this new methodology, only outcomes that are relevant in determining preference over a given pair of outcomes are considered. Using this new approach, partial COP-nets can be constructed dynamically or on demand as opposed to the current process of generating the entire structure. Experimental results show that the new method yields enormous savings in time and memory requirements, and only a modest reduction in prediction accuracy, with one such large example showing only a 5% decrease in the success rate, while reducing computation time from over 3.5 hours to just 2 seconds.

Title: *Pipelined Projection Algorithms for FPGA Technology*
Student: J. John G. S. Cole
Supervisors: Dr. L. E. Garey, Dr. K. B. Kent
Abstract

Linear systems of equations with Toeplitz coefficient matrices often appear in mathematics and applied science problems. Systems of this form often arise as a result of finite difference methods when used to approximate differential equations with boundary conditions. The sparse structure of Toeplitz matrices lend themselves well to iterative algorithms, such as the projection methods, and are highly favored techniques for solving large systems of linear equations. Field Programmable Gate Arrays (FPGAs) have been growing in popularity among the scientific community due to the potential for impressive performance when evaluating floating point operations in parallel [21]. The regular, sparse, structure inherent in Toeplitz systems admits a highly efficient FPGA projection algorithm pipeline. We will analyze a selection of projection algorithms and derive optimal pipelined hardware designs. These high level designs allow us to predict, with a high level of accuracy, the actual performance that is obtainable. The steepest descent algorithm, discussed later in this paper, was chosen from among the available projection algorithms. The design for the steepest descent pipeline was prototyped on an Altera DE2 FPGA development board. The actual results, matching the theoretical predictions, compare favorably against an efficient software implementation of the algorithm. Development time was a considerable drawback during development, yet many reusable components were produced which could be used to implement various other FGPA algorithms.

Title: *Assigning Routes and Wavelengths for Collaboration over Optical Networks*
Student: Yosri Harzallah
Supervisors: Dr. Bruce Spencer, Dr. Joseph Horton, PhD, Faculty of Computer Science
Abstract

The Routing and Wavelength Assignment (RWA) problem is: Given the topology of a wavelength-routed optical network and a set of traffic demands, what is the optimal route and wavelength for each connection? In this thesis, the focus is on solving the RWA problem in the specific context of collaboration while taking into account the availability of the shared non-network resources. Unlike the previous studies, a traffic demand is defined as a set of one- to-one simultaneous connections to model the traffic of collaboration over a non-multicast capable optical network. The online and offline versions of the problem are discussed, with and without the time as a parameter for scheduling purposes. The case where the network is meant to carry time-multiplexed traffic over the wavelengths is also discussed. Also, the problem of rescheduling a blocked demand is studied. Several solutions based on Integer Linear Programs (ILP) and heuristics are pro-posed, implemented and their performance compared. The offline case is solved using two types of ILPs: link and path formulations. ILPs are also proposed for the online problem in addition to the heuristic algorithms called SLF, MCF and SSF. While the link formulations give optimal solutions, they take a long time to solve and thus they can only be used for small problems. Path formulations and heuristics scale better but at the expense of optimality. The online approach is recommended when the resort to an offline approach is forbidden by the size of the problem.

Title: *Design and Implementation of a Phishing Filter for Email Systems*
Student: Luyin Huang
Supervisor: Dr. Weichang Du
Abstract:

Phishing attacks are becoming a very common way to pirate information on the Internet and are also the hardest to trace. The goal of this new and recent method of hacking is very dangerous because it is one of the simplest as it depends only on user trust. The objective is to cheat a user into divulging personal information by leading them to an unsafe website that looks genuine and similar to the original website the user may be used to visiting. This leads to identity theft and the exploitation of vulnerabilities from the fact that most users will not second guess the authenticity of the link they use. This report focuses mainly on keeping users aware of the links they use. It also aims on creating a filter to analyze message content for information integrity in order to assure that potentially unsafe websites are clearly noted in the message before the user mistakenly assumes the opposite.

Title: *AUTOMATIC RULE TUNING IN INTRUSION DETECTION SYSTEMS: ONLINE AND OFFLINE*
Student: Shah Arif Iqbal
Supervisor: Dr. Ali A. Ghorbani
Abstract

With the frequent changes in network environments, managing and updating the rule based system has become a very challenging task for the administrator. Usually, rule based systems work to make sense of a huge amount of alerts generated by the intrusion detection systems (IDSs) every minute. Therefore, it is very important to make sure that these systems are error-free and that the rules are appropriate for the current network. This issue is addressed by Rule Tuning, which automatically tunes the rules based on the current network environment. Several rule tuning methods have been proposed in the literature, but none of them are explicitly interested in keeping the structure of the rules intact. However, it is crucial to keep the structure intact because the rule structures are created based on expert knowledge and should only be allowed to be modified by the experts. Thus, the problem with the rule tuning is to tune the internal thresholds and to keep the structure intact. In this thesis, we propose two methods for tuning the rules, online and offline. Both the methods do the threshold tuning without modifying the structure of the rules. Here, our approach for online threshold tuning is to monitor the alerts and detect steady changes in them. And then, based on the changes we detect in the generated alerts we tune the appropriate thresholds. Again, for offline tuning, we have set a target number of firing for a rule-set and we tune the thresholds to achieve the target. We have implemented both the methods and evaluated them using real-world datasets collected by SNORT. Our approaches were successfully able to tune the rules in all the cases with marginal error.

Title: *A Learning-based Multi-document Summarization Framework*
Student: Mohammad Amin Jashki
Supervisor: Dr. Ali Ghorbani
Abstract

Multi-document summarization is an automatic process of extracting information from multiple documents compiled about the same topic. The resulting product helps the readers to familiarize themselves with the information contained in that group of text documents. In this thesis, a new framework for summarizing multiple documents is proposed. The core of the framework is a Conditional Random Field learning model which is based on several inter-sentence text mining features. Furthermore, the framework is equipped with a novel feature selection algorithm to cluster text documents. The clustering algorithm is utilized to find documents on the same topic. Several experiments have been conducted to analyze the performance of the framework. The experimental results show that the feature selection algorithm is successfully selecting appropriate features for document clustering. The results also confirm that the summarizer framework is statistically competitive in producing summaries.

Title: *EVALUATING THE APPROPRIATENESS OF SPEECH INPUT IN MARINE APPLICATIONS*
Author: Nathan Langton
Supervisors: Dr. Joanna Lumsden, Dr. Jane Fritz, Dr. Irina Kondratova
Abstract

As mobile Human Computer Interaction matures as a discipline, a number of novel evaluation approaches for mobile technologies are emerging. Currently, however, there is no generally agreed consensus on how best to evaluate mobile technologies. The benefit in terms of validity and usefulness of lab-based versus field-based evaluations of mobile applications is a subject of considerable ongoing debate. The infancy of the discipline means that there is currently

relatively little literature or empirical data available on the effect of evaluation environment on the results obtained during empirical assessments. As such, this debate is often viewed as a matter of opinion. The research presented in this thesis document had two objectives: (1) to determine the efficacy of speech as an accurate data-input mechanism for mobile applications in marine environments (specifically, lobster fishing vessels); and (2) to investigate the effect of evaluation environment on results obtained during empirical assessments. This thesis document describes a tri-study comparison of field and lab-based evaluation approaches within our complex context of use. We demonstrate that it is possible to conduct a meaningful, and thus a contextually relevant, lab-based evaluation of user performance of a mobile system designed to be used within a contextually rich and complex, real-world environment. Furthermore, our results strongly support the potential for effective use of speech for data collection and vessel control onboard lobster fishing vessels.

Title: *Ontology Validation under the Closed-World Semantics*
Student: Cheng Lu
Supervisors: Dr. Bruce Spencer, Dr. Weichang Du
Abstract

In today's business world, large quantity of data is usually stored in the database format for processing. In recent years, Knowledge Base (KB) represented in the ontology format has been developed as an alternative way to store data. However, the common validation operation applied to KB does not always perform as expected from a closed- world perspective. Knowledge Base typically is supposed to represent an \open-world". For the relational database theory, the database domain is always a \closed-world". This difference is often referred to as “open-world" vs. “closed-world". Most common Description Logic (DL) reasoners assume a KB is in open-world when performing reasoning tasks on it. The results from these reasoning tasks do not always satisfy the users who view the KB with a database perspective which typically is closed-world. In this thesis, we propose an approach validating a KB under the closed-world semantics. We design and implement a DL ontology reasoner prototype which is capable of dealing with both open-world reasoning tasks and closed-world reasoning tasks. Reasoning with a KB that is partially closed using the `K' operator is also discussed in this thesis. Traditional database users will have a flexible way to express that some parts of the KB are open and some are closed by using K-operator reasoning services.

Title: *SELF-ORGANIZING STRUCTURED NETWORKS AND AGENT-BASED SOFTWARE DESIGN*
Student: Adam MacDonald
Supervisor: Dr. Mihaela Ulieru
Abstract

The design and implementation of an agent-based model for self-organized networks is presented. Localized agent-level rules are specified which result in the construction of an emergent network configuration with specific characteristics. The model can produce networks which represent a precise geometric structure, exhibit resilience to unexpected events, and respond to constraints in a physical environment. It uses a structural-level coordinate system and localized information passing to allow each agent to be aware of its unique position within the

structure. The design of generic agent-based simulation software is presented as well as solutions to specific technical problems relevant to agent-based simulation.

Title: *Mutual Reinforcement of Word and Document Clusterings: A Parallel Approach*
Student: Majid Makki
Supervisors: Dr. Ali A. Ghorbani, Dr. Hamid R. Rabiee
Abstract

This thesis addresses the problem of dimensionality reduction in document clustering. The ultimate goal of the thesis is to propose a solution in the form of a framework based on the mutual reinforcement of word clustering and document clustering. The idea is to initially cluster documents based on a subset of the original feature set and then expand the feature set using supervised distributional word clustering. Expectation-Maximization (EM) is employed to adopt the method as an unsupervised one. To overcome the high time complexity imposed by EM, parallelization is applied by means of sampling methods and unsupervised ensemble learning. Four concrete versions of the framework are implemented and the behavior of each is studied along with two rivals on three data sets. The upshot of the experiments is that the versions of the framework are comparable to their rivals on the two smaller data sets and better on the largest data set in terms of the trade-off they can make between processing time and accuracy. Moreover, it is shown that the results obtained by all the methods are roughly the same according to an internal evaluation measure.

Title: *Ontology-based Unit Test Generation*
Student: Valeh Hosseinzadeh Nasser
Supervisors: Dr. Weichang Du, Dr. Dawn MacIsaac
Abstract

Various software systems have deferent test requirements. In order to specify adequate levels of testing, coverage criteria are used. Providing a tool for test experts to define custom coverage criteria can potentially increase the quality of test suites generated through automation. This thesis investigates application of knowledge engineering techniques in order to offer such controls to test experts. The method which is presented in this work facilitates the enrichment of test oracles with test experts' mental model of error-prone aspects of software, and definitions of custom coverage criteria. The knowledge that is referred by coverage criteria for test case selection may not be present in standard test oracles, such as UML state machines. To solve this problem an extensible representation of test oracles is required to allow addition of expert knowledge. Also, to enable test experts to add knowledge freely, the test case selection algorithms should be decoupled from the knowledge, which is used in test case selection. For this reason, the test oracles are represented in ontologies, which are highly extensible. The coverage criteria are written in a rule language, using the vocabulary defined by the test oracle ontology. This approach makes it possible for the test experts to add the knowledge to the test oracles and compose new coverage criteria. To decouple the knowledge that is represented in test oracles from the test selection algorithms, reasoning is used for test case selection. Prevalent test

case generation technologies are then used for generating the test cases. The focus of this thesis is on unit testing based on UML state machines.

Title: *SELECTING PAYLOAD FEATURES USING N-GRAM ANALYSIS TO CHARACTERIZE IRC TRAFFIC AND MODEL BEHAVIOUR OF IRC–BASED BOTNETS*
Student: Goaletsa Rammidi
Supervisor: Dr. Ali Ghorbani
Abstract

A botnet is a network of compromised computers remotely controlled by an attacker. Different feature selection methods are applied to find a lower dimension subset of Unicode characters as payload features, using n-gram analysis with n=1, to classify TCP packets into IRC and non-IRC application communities. The identified IRC packets are grouped into 1 minute intervals to create a temporal frequent distribution, and then unsupervised clustering is applied to separate botnet IRC from normal IRC. The botnet cluster is labeled as one with minimum cluster standard deviation. We found a subset of 9 features that separate IRC packets from non-IRC in less time and comparable accuracy to using all the 256 features. We also found that IRC traffic is dominated by the first 128 Unicode characters, therefore, using all the 256 may not be necessary. Clustering packets into IRC and non-IRC using merged Xmeans had lower false alarm rates than Kmeans and consistent high detection rates than unmerged Xmeans.

Title: *RNA Structural Motif Discovery using Probabilistic Tree Adjoining Grammars*
Student: Emad Bahrami Samani
Supervisor: Dr. Patricia Evans
Abstract

Patterns in both coding and non-coding RNAs are indicators of significant biological functions. RNA motif discovery searches for a set of sequences which fold into a structure that performs a specific biological function. Finding novel RNA motifs has great applicability in medicine. This thesis proposes a new technique based on probabilistic tree adjoining grammars (TAG) to solve this problem. The extra power provided by TAGs to describe the crossing dependencies seems to be the right kind of ability for the practical problem of modeling RNA structure including pseudoknots. A probabilistic tree adjoining grammar is used to capture the structural characteristics of RNA molecules. The trained model is then used to parse the structure of the test data. The similarity estimation part calculates the mutual information metric for derivation sequences in the process of parsing. It is shown that the similarity measure proposed in this research is able to capture the relationships between different parts of the RNA structures. The principle coordinate analysis (PCoA) algorithm is used to produce 2-D projection of the distance matrices which are examined via k-means clustering. The main advantage of this method is that it can efficiently model pseudoknots in RNA secondary structures and extract motifs containing pseudoknotted structures. The algorithm is able to successfully identify biological RNA motifs based on the secondary structure data.

Title: *Heterogeneous Parallelization for RNA Structure Comparison*
Student: Eric Snow
Supervisors: Dr. Eric Aubanel, Dr. Patricia Evans
Abstract

This thesis describes the parallelization of a dynamic programming algorithm used to _nd common RNA secondary structures including pseudoknots and similar structures. The sequential algorithm is recursive and uses memo-ization and data-driven selective allocation of the tables, in order to cope with the high space and time demands. These features, in addition to the irregular nature of the data access pattern, present particular challenges to parallelization. A new manager-worker approach is presented, where workers are responsible for task creation and the manager's sole responsibility is overseeing load balancing. Special considerations are given to the management of distributed, dynamic task creation and data structures, along with general inter-process communication and load balancing on a heterogeneous computational platform. Experimental results show an average speedup of 7 using 64 processors, a modest increase, along with a highly scalable level of memory usage. This allows for the comparison of much longer and more complex RNA molecules than is possible in the sequential implementation, with molecules of up to 4000 bases tested.

Title: *Multi-Agent Systems in Simulation: A Study on Recent Research & Illustration through Development of an Urban Traffic Simulation*
Student: Vinaykumar Sukumar
Supervisor: Dr. P.K. Mahanti
Abstract

This research report aims to provide a comprehensive study on recent research and development in multi-agent systems in the field of decision support systems using simulation. As part of the study an urban traffic simulation application is developed to help in deciding the priority based timing for traffic signals in busy intersections. This application is developed using MASON, which is a fast discrete-event multi-agent simulation java based framework. Through the development of this application and the literature survey, this study will illustrate the usefulness of multi-agent simulation in the area of decision support. It highlights the contribution made by researchers in the use of multi-agent systems in simulation and identifies key areas of recent research and recommendations for the future.

Title: *Automated Cryptanalysis: attacking classical ciphers*
Student: Eric Tucker
Supervisor: Prof. Rodney H Cooper
Abstract

Cryptography is the study of developing methods to hide or conceal messages intended to be kept secret from those individuals other than intended receivers. Cryptographers put a great deal of effort into “breaking” the very same systems they develop in order to better understand and invent algorithms for this purpose. A cryptosystem which can be compromised easily to reveal a secret message is not very useful. The focus of this thesis is to investigate both *Monoalphabetic Substitution* and *Transposition* Ciphers, two classical algorithms, using procedures developed for

deployment on a modern personal computer. This thesis explores the extent to which these classical ciphers can be compromised without human intervention during the process.

Title: *Multicasting in Wireless Ad-Hoc Networks*
Student: Doga Tav
Supervisors: Prof. John DeDourek, Dr. Premyslaw Pochec
Abstract

Multicasting has many applications over the Internet, including teleconferencing, broadcasting TV shows, multi-player gaming, and many other such applications. Multicasting uses an underlying network structure to deliver packets efficiently to multiple destinations simultaneously. Without multicasting, a source would have to send the same packet to every node (multicast subscriber) individually. In our research, we observe properties of multicasting for wireless ad-hoc networks in which wireless nodes may enter and leave the multicast group randomly. Wireless ad-hoc networks have a decentralized structure which allows every node to act as a router and a receiver. Packets routed in this type of network are expected to behave differently than in wired networks. Our results show us that Protocol Independent Multicast - Dense Mode (PIM- DM) protocol gives the highest data rate compared to Protocol Independent Multicast - Sparse Mode (PIM-SM) and Distance Vector Multicast Routing Protocol (DVMRP).

Title: *An Efficient Algorithm for 3D Image Registration*
Student: Matthew D. Williamson
Supervisors: Dr. Bradford G. Nickerson, Dr. Tom A. Al, Geology
Abstract

Computed Tomography (CT) scans of rock samples can yield high resolution three dimensional (3D) images. Because a sample is imaged over time, the result is a set of 3D images (in our case 3D 16-bit gray scale images of size 10243). Since it is unlikely that the sample has been placed into the CT scanning device in the exact same orientation each and every time an image is acquired, a need for image registration has been identified. We present a new algorithm for two 3D images that grows features, computes the transformations from one image such that interesting features are transformed to the same coordinate system as the other, and finally resamples one of the images to align (register) it precisely with the second image. We compare our 3D image registration algorithm to a commonly used ImageJ plugin called TurboReg using 8 time series 3D images of sandstone and 7 of dolomite. This algorithm is based on a pyramid approach to subpixel registration. Our results show an average difference improvement of 19.93% and 2.02% for the sandstone and dolomite images, respectively, when testing against Turboreg. Inside the cylindrical region of interest, we found the ratio of average difference (standard deviation) between the resampled images (compared to a chosen reference image) and the original images (again compared to the chosen reference image) to be 0.82 (0.49), respectively, for the 9 sandstone images, and 0.53 (0.91) for the 8 dolomite images.