Apr 17th, 12:00 AM

Proceedings, MSVSCC 2014

Old Dominion University, Department of Modeling, Simulation & Visualization Engineering

Old Dominion University, Virginia Modeling, Analysis & Simulation Center

OLD DOMINION UNIVERSITY
IDEA FUSION

Virginia Modeling, Analysis and Simulation Center

Modeling, Simulation & Visualization
2014 Student Capstone Conference

Final Proceedings

April 17, 2014
Old Dominion University

2014 Student Capstone Conference Final Proceedings Introduction

2014 marks the eighth year of the VMASC Capstone Conference for Modeling, Simulation and Gaming. This year our conference attracted a number of fine student written papers and presentations, resulting in 32 research works that were presented on April 17, 2014 at the conference.

The tracks that the papers were divided up into included the following:

- General Sciences Application
- Business, Industry, Infrastructure Security, & Military Application
- Medical & Healthcare Application
- Gaming & Virtual Reality Application
- Agent-Based M&S

For each track there were two awards given out- The Best Paper award, and the Best Presentation Award.

Overall Best Paper- The Gene Newman Award

The overall best paper is awarded the Gene Newman award. This award was established by Mike McGinnis in 2007; the award is presented to the outstanding student for overall best presentation, best paper, and research contribution. The Gene Newman Award for Excellence in M&S Research is an award that honors Mr. Eugene Newman for his pioneering effort in supporting and advancing modeling and simulation. Mr. Newman played a significant role in the creation of VMASC by realizing the need for credentialed experts in the M&S workforce, both military and industry. His foresight has affected both the economic development and the high level of expertise in the M&S community of Hampton Roads. The Students receiving this award will have proven themselves to be outstanding researchers and practitioners of modeling and simulation.

The following proceedings document is organized into chapters for each of the tracks, in the above order. At the beginning of each section is a front page piece giving the names of the papers and the authors.

Medical & Healthcare Application 100

General Sciences Application

VMASC Track Chair: Dr. Yiannis Papelis
MSVE Track Chair: Dr. Masha Sosonkina

Numerical Simulation of Surface Charge Properties for Silica Nanoparticles
Author(s): Selcuk Atalay, Ali Beskok, and Shizhi Qian

Using Eye and Head Movements as a Control Mechanism for Tele-operating a Ground Based Robot and its Payload
Author(s): Kathryn Catlett, Dr. Yiannis Papelis, Dr. Ginger Watson, Dr. James Bliss, and Dr. John Sokolowski

High-fidelity Simulations of Electron Accelerators Using an Innovative GPU-Optimized Code
Author(s): Kamesh Arumugam, Alexander Godunov, Desh Ranjan, Balsa Terzic, and Mohammad Zubair

Modeling Protein Structures Features from Three Dimensional Cyro-EM Images
Author(s): Dong Si and Dr.Jing He

Links Between Operations Research and Modeling & Simulation
Author(s): Daniele Vernon-Bido, and Dr. Andrew Collins

High-Order and Attributed Motifs in Complex Networks
Author(s): David Wright

GPU Accelerated Randomized Singular Value Decomposition and its Application in Image Compression
Author(s): Hao Ji and Dr.Yaohang Li

Intrinsically Disorder Protein Prediction using Undersampling Feedforward Neural Networks and Predicted Amino Acid Features
Author(s): Qiaoyi Li, Steven Pascal, and Dr.Yaohang Li

Thread Affinity, Power, Energy, and Performance on the Intel Xeon Phi
Author(s): Gary Lawson, and Dr.Masha Sosonkina

Using Eye and Head Movements as a Control Mechanism for Tele-operating a Ground Based Robot and its Payload

Kathryn Catlett

Graduate Student
Dept. Of Modeling Simulation & Visualization Engr.
kcatl001@odu.edu

Yiannis Papelis

Research Professor & Adjunct Associate Professor
Virginia Modeling Analysis & Simulation Ctr.
Dept. Of Modeling Simulation & Visualization Engr.
ypapelis@odu.edu

Ginger S. Watson

Assoc. Professor & Interim Chair
STEM Education & Professional Studies
gswatson@odu.edu

James Bliss

Professor & Chair
Department of Psychology
jbliss@odu.edu

John Sokolowski

Executive Director & Associate Professor
Virginia Modeling Analysis & Simulation Ctr.
Dept. Of Modeling Simulation & Visualization Engr.
jsokolow@odu.edu

Old Dominion University

Keywords: Eye Tracking, Human-Robot Interaction, Tele-operated Robotics, Control.

Abstract

To date, eye tracking has been used to study user's attention patterns while performing a task or as an aide for disabled persons to allow hands-free interaction with a computer. However, the increasing accuracy and reduced cost of eye- and head-tracking equipment makes it feasible to utilize this technology for explicit control tasks, especially in cases there is confluence between the visual task and control. The goal of this research is to investigate the use of eye tracking to design a more natural interface for the control of a camera-equipped, remotely operated robot in tasks that require the operator to simultaneously guide the robot as well as perform a visual search around the vehicle through the use of a Pan/Tilt (PT) camera. Three possible methods of control will be investigated: using traditional hand-held controllers, using a hybrid approach that uses one hand-held controller and eye-tracking for payload control and a hands-free approach that only depends on eye/head tracking for controlling the robot and its payload. This paper provides a review of existing related literature and outlines a research study that will allow a performance-based evaluation of the three approaches.

1. INTRODUCTION

The purpose of this study is to determine the feasibility of using intentional eye movement as a control mechanism for tele-operating an unmanned vehicle. While robots are commonly manipulated with interfaces such as two-hand controllers or joysticks, some more sophisticated interfaces may require the use of both hands and feet. Thus, this project aims to investigate the use of eye movement as another layer of user interaction and control.

Human beings naturally perform constant visual scans of their environment [1]. When performing a visual search, the human eye moves rapidly between fixations, only dwelling on a single point for about 200 to 400 msec. The length of time for a person scanning a scene is typically about 5 degrees of visual angle, and for eye movements that cover more than 20 degrees, the head moves as well [2]. This project aims to take advantages of these visual search patterns to design a more natural interface for the control of the robot and camera.

2. CURRENT LITERATURE

Eye tracking is currently used in a number of research applications, usually split between two categories: diagnostic and interactive. Diagnostic applications use the eye tracker to provide objective or quantitative feedback concerning the user's attention or reactions to stimulation. Interactive applications use the eye tracker as another input device, either replacing the usual mouse and allowing the eyes to control the location of a cursor on a screen or using knowledge of the user's gaze to alter what is seen on the display [3].

Control theories primarily focus on relying on eye tracking for disabled persons; however Manu Kumar et al proposed the use of eye tracking to enter ATM passwords in order to reduce the chance of the password being stolen by a "shoulder surfer" [4]. Their study touched factors that may

attribute to the effectiveness of using eye tracking, such as the distance between keys on the screen and modes of input: dwell or trigger. They concluded that the speed difference between the dwell and trigger methods was inconclusive. The trigger approach, however, showed significantly higher error rates due to users having difficulty properly timing their hand and eyes to coordinate perfectly [4].

Such an approach to security is inspired by the systems that allow disabled persons to interact with a computer using just their eyes. One such system, called the Erica project, has been detailed by Thomas E. Hutchinson et al [5]. This project uses a tier fixation system for selection. When the user's eyes have fixed for two to three seconds on a certain point, Erica plays a sound and a cursor appears in line with the gaze. If the user continues to focus on this cursor, a second tone sounds and the point on which they are focused is selected.

When it comes to robotics, eye tracking is commonly used in the diagnostic sense. Raj M. Ratwani et al used eye tracking to study situational awareness while a single user attempted to control several robots. The user was required to direct five semi-autonomous UAVs to specific targets on a map while avoiding dynamically moving hazard areas. The eye tracker was used to monitor the user's attention in order to measure the user's cognitive processing and predict his situational awareness. This was done by measuring how much time a user focused on a specific robot or task as well as tracking his scan patterns [6].

Such scan patterns have also been used to study the effectiveness of graphical user interfaces for UAVs. By determining where the pilot of the UAV tends to focus, a user interface can be streamlined to make sure information is easily available. Tvaryanas performed a study that required users to perform certain actions with the UAV (turning a certain number of degrees, maintaining a certain altitude, etc.). He tested how quickly they could make those changes and how well they could maintain them while tracking their gaze. This proved the necessity of an ergonomic user interface layout for the effective piloting of a UAV [7].

3. EYE/HEAD TRACKING

This project will make use of eye and head tracking not as a diagnostic tool, but as a possible control mechanism for the remote control of an unmanned ground vehicle and a camera payload. In order to allow the eye and head tracking to serve as a controller, however, the method of control and measurement must be determined. There are a number of options for each that would alter the effectiveness and ease of use of the system.

3.1. Control Schemes

One way to classify control inputs when tele-operating a robot is by separating them into analog and binary control commands. Analog control commands provide a continuous signal that is used to control an aspect of the robot across a range. A binary control command only has two values, on or off, and is used to control an aspect of the robot that can only have two states.

Table 1 General control schemes by input type

	Joysticks	**Mouse**	**Eye Tracker**
Binary	Button press	Click	Blink, Nod, Fixation Time
Analog	Deflection	Cursor distance from reference pt.	Fixation distance from reference pt.

When using a joystick, an analog control would be the deflection of the joystick, whereas pressing a button or flipping a toggle switch would be an example of a binary control. Similarly, when using a mouse, an analog control can be derived by measuring the distance that the mouse has traveled from a reference point, where a mouse click provides a binary control. Note that this classification only refers to the control inputs, not the effect on the robot. For example, a binary control can be used to control the speed of the robot simply by setting the speed to a pre-determined value; however it is more natural to pair each control type to a matching effect on the robot.

In the case of eye tracking, analog control can be derived similar to a mouse, i.e., by establishing a reference point and then measuring the distance to the current eye fixation location. A fixation is determined by calculating the standard deviation of the eye position over a time window, and if that standard deviation is less than a determined measurement, the centroid of the user's focal area is considered the fixation point. There are several options for deriving a binary control. These include blinking, blinking twice, nodding of the head, or fixating at a specific location for an extended time. **Table 1** summarizes this control classification and how it applies to different input devices.

3.2. Hands-Free Binary Control

One of the goals of the project is to identify beneficial control schemes that leverage eye- and head-tracking to replace traditional binary and analog inputs. Binary inputs in particular are important because they are central to selections and making action commitments. Therefore, the project will consider four possible options for hands-free selection.

The first method will be by identifying signature blinking patterns, such as a blink that lasts longer than normal or a double-blink. When such a deliberate blink is detected, the simulation will either select the point at which the user is looking at the time or execute some other command associated with the blink.

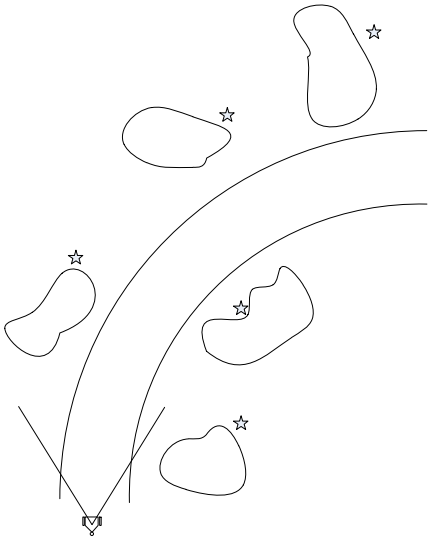


Figure 1 An example leg of the course, in which objects of interest (stars) are hidden from the forward view of the robot

The second makes use of the fact that the eye tracker estimates the position and orientation of the head in order to track the eyes. This means that if the user makes a sharp head movement, the eye tracker will pick it up and can react accordingly.

The third selection method would be to use fixation. If the user focuses on a specific point for a certain time, then this will be considered a binary control applied to a context-specific selection. In a similar manner to the Erica project, this method will display some visual feedback to ensure that the user wishes to select the focal point. This may be a shrinking circle around the fixation that, once the circle is small enough, changes color and then selects the point. This way, the user may deliberately make or abort a selection.

The final selection method would be to use a composite binary input, for example fixation plus another binary input, such as a blink. Thus, if the user is focused on a point for a certain amount of time, the simulation would then watch for the signature blink, nod, or other input.

4. SIMULATION SETUP

In order to test the feasibility of eye tracking as a control mechanism for a robot and PT camera, a simulated robot will be placed into a virtual world consisting of a flat terrain populated with a variety of obstacles and objects of interest. Obstacles may include buildings, cars, trees, etc., whereas objects of interest will be items such as IEDs, disabled individuals, etc.

The purpose of the task will be for the operator to follow a route through the environment and search for objects of interest. There will be a desired course laid out for the user to follow, comprised of a mixture of curves and straightaways. This path, which may be considered a "safe zone" for an operation, will be roughly five times the width of the robot. The user must stay within this course during his search. A variety of objects of interest will be placed in the world in such a way that they are visible only if the user pans and/or tilts the PT camera. This is to ensure that simply monitoring the front-viewing drive camera is not enough to see the objects of interest. An example of this is shown in **Figure 1**.

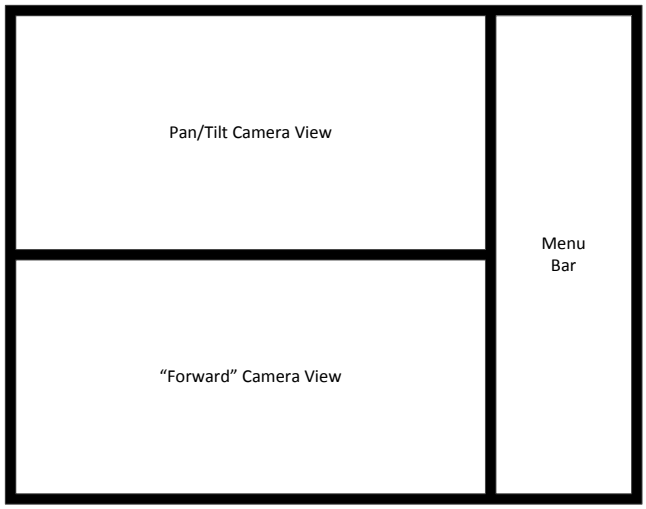


Figure 2 The proposed layout of the user interface

The user will view this world through a split-screen interface with three distinct areas of interest, as seen in **Figure 2**. On the right hand side will be a menu bar stretching the full height of the screen. This bar will hold various buttons to allow for the control of the robot's speed as well as to signal when the user believes he has found a target.

The remainder of the screen will be split horizontally into two views of the world. The top view will be a camera with pan and tilt functionality (the PT camera). This will be the view used to scan the environment for target objects. The bottom view will be the drive camera, mounted on the front of the robot and thus always pointing "forward."

5. CONTROL MECHANISMS

Three possible methods of robot tele-operation will be compared:

1) Manual: two joysticks, one to control the velocity and direction of the robot and one to control the pan and tilt angle of the camera.
2) Hybrid: a joystick for the movement of the robot and eye tracking for the movement of the camera.
3) Hands-Free: eye tracking for the movement of both the robot and the camera.

5.1. Manual

The first method uses two hand-held controllers or joysticks, one to control the movement of the robot and one

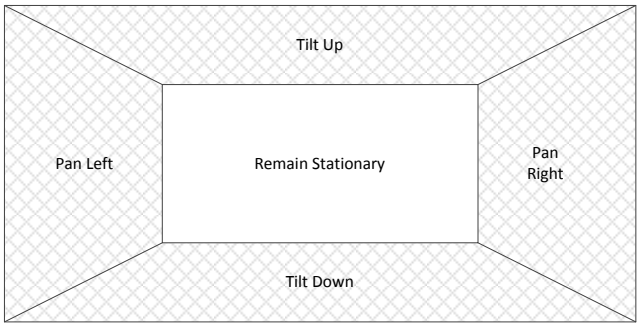


Figure 3 The areas of interest in the PT camera view

to control the movement of the camera. Tilting the movement joystick forward will cause the robot to move forward at a speed proportional to the deflection of the joystick. Tilting the joystick left or right will turn the robot towards that direction. Letting go of the joystick will cause the robot to come to a stop.

Tilting the camera joystick forward or backward will cause the PT camera to tilt up or down. Likewise, tilting the camera joystick left or right will cause the PT camera to pan in that direction. Releasing the camera joystick will cause the PT camera to stop moving. This method of control is often used in video games to control both the movement of a character and the direction it is pointing. This method can be considered the traditional control technique. Analog controls are mapped to analog behaviors of the tele-operated robot, and binary controls such as buttons can be used to make discrete selections like signaling the identification of an object of interest.

5.2. Hybrid

The second method of controlling the robot uses a mixture of hands and eyes. The joystick in this method controls only the movement of the robot, whereas the eyes are used to control the PT camera. The fixation point will be used as an analog signal to control the velocity of the pan and tilt of the camera. The software will continuously evaluate fixation and if the fixation is located within four screen areas, the distance of the fixation to the edge of the area will be treated as an analog control whose value will be applied to the respective speed of the pan and tilt. The concept is illustrated in **Figure 3**.

Fixating on the upper portion of the screen will cause the camera to tilt up at a speed proportional to the vertical distance between the fixation point and the bottom edge of the 'Tilt-Up' area. Similarly, a fixation inside the 'Pan Right' area will be used to pan the camera to the right at a speed proportional to the distance between the fixation and the left edge of the Pan Right area. If no fixation is detected, or if the fixation is outside the shaded areas, the camera will stop. According to [1], when performing a visual search there are no fixations; however, once a point of interest has been identified, focusing on it will create a fixation, and as

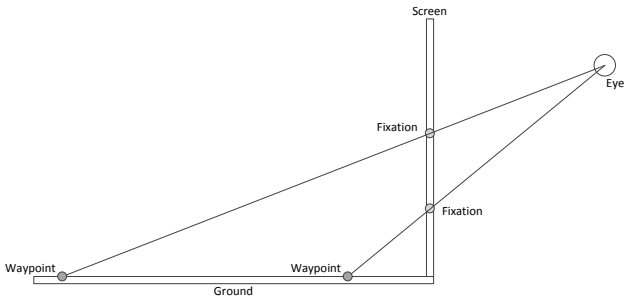


Figure 4 Mapping the screen fixation point to a waypoint in the world

the robot moves, the object will tend to move to the edge of the screen, and as long as the user fixates on the object, the camera will automatically pan to keep the object within view.

5.3. Hands-Free

The third and final control method uses eye tracking for both guidance and visual search. The visual search approach will be identical to the Hybrid method, and will be active only when the user's fixation is within the PT view. Controlling the robot involves two components, speed and direction.

The current plan is to utilize discrete selections to control the robot's speed by selecting among a fixed set of per-specified speeds (stopped, low, medium, fast). To do so, the user will fixation within the appropriate button on the right side of the screen, and once committed, the longitudinal speed of the robot will be set to the appropriate value.

Controlling direction will be done by allowing the user to designate a way-point or a target destination simply by looking at that point in the drive camera view. By knowing the location of the eye and the fixation point on the screen, the system will calculate the angle of the gaze, as illustrated in **Figure 4**. With this gaze trajectory, the fixation point can be projected onto the flat surface of the virtual world. The robot will then trace a curved path to this waypoint.

If the user has not provided a new waypoint by the time the robot has reached its current target, then the robot will continue forward at the provided speed. This will require the user to carefully split his attention between the PT camera and controlling the movement of the robot, else the robot will travel outside of the designated "safe zone." The time spent outside of the "safe zone" will be measured for each run of the simulation and will be used as one of the performance measures of the control mechanisms. **Table 2** summarizes robot operation for each control method.

Table 2 Specific input methods by control mechanism

	Joysticks	**Mixture**	**Eyes Only**
Stop Moving	No Deflection	No Deflection	Blink at Button
Speed Control	Forward Deflection	Forward Deflection	Blink at Button
Steering	Sideways Deflection	Sideways Deflection	Fixation Time
Camera Control	Deflection	Distance from Center Point	Distance from Center Point
Signal Target Found	Button Press	Button Press	Blink at Button

6. EXPERIMENTAL DESIGN AND PERFORMANCE MEASURES

The exact conditions that will be tested are still under consideration. The plan is to recruit students for participation in the study. Students will be briefed as to the goal, and given a practice session as necessary. They will then be allowed to guide the robot while looking for objects of interest.

In all cases, the minimum set of performance metrics will include: the speed at which the user completes the task, the accuracy by which the user stays within the designated area measured as the amount of time the robot spends outside the 'safe' area, and the number of targets the user locates. The users will also be filling out a questionnaire before and after the experiment in order to gain some feedback concerning any prior experience using the control methods as well as their impressions of each.

7. CONCLUSION

Overall, this proposed research will test the practicality of eye tracking as it applies to the control of robots. While eye tracking has proven useful as a selection tool and performance measure, its usefulness for more complex actions remains to be seen.

Results of this research may vary based on the level of control that the eye tracking maintains. Using the eye tracker to select waypoints may prove too time consuming for the user depending on the method of selection used. That level of control also requires the most splitting of attention and runs the risk of producing the most variance from the designated path. While the combined method is expected to be the most efficient, the conventional method may prove to still be the best, since it is familiar to most users.

References

[1] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. London: Springer, 2003. Print

[2] C. Ware. *Information Visualization: Perception for Design*. 3rd Edition. Amsterdam: Morgan Kaufmann, 2013. Print.

[3] A. T. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4), p. 455-470 (2002).

[4] Kumar, M., Garfinkel, T., Boneh, D., and Winograd, T. Reducing shoulder surfing by using gaze-based password entry. *Proceedings of the 3rd symposium on Usable privacy and security*, July 18-20, 2007, Pittsburgh, Pennsylvania.

[5 T. E. Hutchinson, K. P. White, W. N. Martin, K. C. Reichert, and L. A. Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics*, 1989.19(Nov.Dec): p1527-1533.

[6] R. M. Ratwani, J. M. McCurry, and J. G. Trafton. Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness. *5th ACM/1EEE International Conference on Human-Robot Interaction*, 2010(March): p243-250.

[7] A. P. Tvaryanas. Visual scan patterns during simulated control of an uninhabited aerial vehicle (UAV). *Aviation, Space, and Environmental Medicine*. 75(6), p. 531-538 (June 2004).

Biography

Kathryn Catlett is a Masters student in the Modeling and Simulation Department of Old Dominion University. She earned her Bachelor's Degree in Computer Science from Auburn University. Kathryn currently works as a research assistant at the Virginia Modeling, Analysis, and Simulation Center under the VMASC Student Scholar program.

High-fidelity simulations of electron accelerators using an innovative GPU-optimized code

Kamesh Arumugam[1,2], Alexander Godunov[3,2], Desh Ranjan[1,2], Balša Terzić[4,3,2] and Mohammad Zubair[1,2]
[1]*Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529*
[2]*Center for Accelerator Science, Old Dominion University, Norfolk, Virginia 23529*
[3]*Department of Physics, Old Dominion University, Norfolk, Virginia 23529*
[4]*Center for Advanced Studies of Accelerators, Jefferson Lab, Newport News, Virginia 23606*

Abstract

Recent development in Graphics Processing Units (GPUs) has enabled a new possibility for highly efficient parallel computing in science and engineering. The advent of multi-core CPUs with a support of multiple GPUs in a cluster has ensured the scalability of the general purpose computing on GPUs. Simulation of coherent synchrotron radiation (CSR) in electron accelerators which involve fast and accurate multidimensional numerical integrations of functions requires a high-performance implementation. In this paper, we present a two-phase algorithm for solving adaptive multidimensional integration on a cluster of compute nodes with multiple GPU devices per node. The algorithm was implemented on a cluster of Intel® Xeon® CPU X5650 compute nodes with 4 Tesla M2090 GPU devices per node. We observed a speedup of up to 240 on a single GPU device and on a cluster of 6 nodes (24 GPU devices) we were able to obtain a speedup of up to 3250. All speedups here are with reference to the sequential implementation running on the compute node. We intend to extend the two-phase algorithm to fit in a CSR model and furthermore develop a highly efficient simulation model for CSR.

Keywords: Adaptive Multidimensional Integration, GPUs, GPU cluster, CSR

1 INTRODUCTION

The development and optimization of new designs of accelerators-based light sources and other electron accelerator machines depends crucially on accurate, high-resolution numerical simulations. As the brightness and energy of accelerator-based light sources is extended beyond present levels, it becomes necessary to improve existing computational modeling capabilities or develop new ones. One of the most critical needs is to develop efficient codes for simulating collective effects that severely degrade beam quality, such as coherent synchrotron radiation (CSR) and CSR-driven microbunching instability. Accurate CSR simulations are needed in: (i) determining the feasibility of light sources; (ii) optimizing light source designs and construction costs; (iii) the design of accelerators with ultra-low emittances and ultra-short beam bunches; and (iv) generating wanted CSR for THz light sources. Operating electron accelerators – electron-based light sources and storage rings and electron colliders – costs hundreds of thousands of dollars per hour. High-fidelity simulations that are able to capture the underlying physics of these machines, which our new code aims to deliver, will not only defray the operation cost but also contribute to fine-tuning the parameters for more efficient operation.

Direct simulation of CSR in 2-D and 3-D is prohibitively costly in terms of efficiency and memory requirements. Consequently, the present CSR codes employ a number of approximations and simplifications that are often inadequate for resolving essential physics in many realistic situations. These situations where existing CSR codes fail are expected to become commonplace as the design of next-generation light sources commences. This provides a strong impetus for the development of the new CSR codes that are both accurate and efficient.

In practice, accurate and high-resolution numerical simulation of CSR has always been limited by the computational requirements of several numerical algorithms in the CSR code, such as adaptive integration, multidimensional interpolation, and charge deposition. The 2-D CSR model developed by Li [16, 17] is based on computation of the retarded potentials by direct integration over a 2-D charge distribution (no vertical size), represented by macroparticles of finite size. Initial profiling of this CSR code suggests that the adaptive integration component used in computing the retarded potential takes 95-99.9% of the overall simulation time. These simulations are very time-consuming on a single-processor system (on the order of months or years) and need to be implemented on the massively parallel computer architectures to reduce the simulation time to reasonable values (on the order of hours or days). This provides a strong motivation for the development of high-performance algorithms for fast and accurate multidimensional numerical integration. Many numerical algorithms have been developed, and are part of standard numerical libraries such as NAG, IMSL, QUADPACK, CUBA and others [12, 13, 19, 22]. Providing reliable estimate for the integral at higher dimension requires considerable amount of CPU time, and often this has to be done with efficient paral-

lel algorithms. However, only a few deterministic parallel algorithms have been developed for adaptive multidimensional numerical integration [4, 5, 14]. Some of the existing parallel algorithms are a simple extensions of their sequential counterparts, utilizing the multithreading nature of the multicore CPU platform and resulting in a moderate speed-up.

Recent emergence of multi-core architectures with CUDA-enabled GPUs, provides a great opportunity and a formidable challenge on the adaptive multidimensional integration front. The sequential adaptive integration algorithms seek to minimize the number of integrand evaluations because this directly corresponds to minimizing the execution time. In contrast, an efficient GPU algorithm must optimize many different components: load balancing, global and local communication, memory management, utilization of registers and cores, etc. This presents a major challenge in developing GPU-optimized algorithms for adaptive multidimensional integration on commonly available and inexpensive hardware.

In this paper, we propose a two-phase algorithm for solving adaptive multidimensional integration on CUDA-enabled GPU platforms. Furthermore we propose a technique to scale this algorithm to multiple GPU devices. The algorithm was implemented on a cluster of Intel® Xeon® CPU X5650 compute nodes with 4 Tesla M2090 GPU devices per node using OpenMP and Message Passing Interface (MPI). We observed a speedup of up to 240 on a single GPU device and on a cluster of 6 nodes (24 GPU devices) we were able to obtain a speedup of up to 3250. All speedups here are with reference to the sequential implementation running on the compute node.

The remainder of the paper is organized as follows. In Section 2, we briefly overview the deterministic methods for adaptive integration and the general equations that numerical CSR simulations are solving. The new parallel algorithm and its implementation for GPU architecture is presented in Section 3. In Section 4, we apply the new parallel algorithm to a battery of functions and discuss its performance. Finally, in Section 5, we discuss our findings and outline the future work.

2 BACKGROUND

In this section, we provide an overview of the general equations that numerical CSR simulations are solving. We then briefly overview deterministic methods for adaptive integration.

2.1 Coherent Synchrotron Radiation

CSR is an effect of curvature-induced self-interaction of a microbunch with a high charge as it traverses a curved trajectory. It can cause a significant emittance degradation, as well as fragmentation and microbunching of the beam bunch. The dynamics of an electron in the beam bunch is captured by the Vlasov-Maxwell equations:

$$\partial_t f + \upsilon \cdot \boldsymbol{\nabla} f + e(\boldsymbol{E} + \boldsymbol{\beta} \times \boldsymbol{B}) \cdot \boldsymbol{\nabla}_{\mathbf{p}} f = 0, \tag{1}$$

$$\upsilon(\boldsymbol{p}) = \frac{\boldsymbol{p}/m_e}{\sqrt{1 + \boldsymbol{p} \cdot \boldsymbol{p}/(m_e c)^2}}, \tag{2}$$

$$\boldsymbol{E} = -\boldsymbol{\nabla}\phi - \frac{1}{c}\partial_t \boldsymbol{A},\ \boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A} \tag{3}$$

$$\begin{bmatrix}\rho(\boldsymbol{r}, t) \\ \boldsymbol{J}(\boldsymbol{r}, t)\end{bmatrix} = \int_0^\infty \begin{bmatrix}1 \\ \boldsymbol{\upsilon}(\boldsymbol{p}, t)\end{bmatrix} f(\boldsymbol{r}, \boldsymbol{p}, t) d\boldsymbol{p}, \tag{4}$$

$$\begin{bmatrix}\phi(\boldsymbol{r}, t) \\ \boldsymbol{A}(\boldsymbol{r}, t)\end{bmatrix} = \int_0^\infty \begin{bmatrix}\rho(\boldsymbol{r}, t) \\ \boldsymbol{J}(\boldsymbol{r}, t)\end{bmatrix} \left(r', t - \frac{||\boldsymbol{r} - \boldsymbol{r}'||}{c}\right) \frac{dr'}{||r - r'||} \tag{5}$$

where $\boldsymbol{p}$ is the particle momentum and $f \equiv f(\boldsymbol{r}, \boldsymbol{p}, t)$ is the distribution function (DF) in phase space $\Gamma = (\boldsymbol{r}, \boldsymbol{p})$, and t' is the retarded time, defined as $t' = t - |\boldsymbol{r} - \boldsymbol{r}'|/c$. m_e is electron mass , and c the speed of light. Also, $\boldsymbol{\beta} \equiv \boldsymbol{\upsilon}/c$, $\boldsymbol{E} = \boldsymbol{E}^{\text{ext}} + \boldsymbol{E}^{\text{self}}$, $\boldsymbol{B} = \boldsymbol{B}^{\text{ext}} + \boldsymbol{B}^{\text{self}}$. Here $\boldsymbol{E}^{\text{ext}}$ and $\boldsymbol{B}^{\text{ext}}$ are external electromagnetic (EM) fields fixed by the accelerator lattice, and $\boldsymbol{E}^{\text{self}}$ and $\boldsymbol{B}^{\text{self}}$ are the EM fields from the bunch self-interaction, which depend on the history of the bunch charge distribution ρ and current density $\boldsymbol{J}$ via the scalar and vector potential ρ and $\boldsymbol{A}$.

As can be seen from the equations above, computation of the potentials requires integration over the history of the charge distribution and current density. The equations point to the main computational bottlenecks of the CSR simulations: (i) data storage for the time-dependent beam quantities (ρ and $\boldsymbol{J}$); (ii) numerical treatment of retardation and singularity in the integral equation for retarded potentials; and (iii) accurate and efficient multidimensional integration in the equation for retarded potentials.

2.2 Adaptive Integration Methods

Adaptive integration is a recursive technique in which a quadrature rule is applied on an integration region to compute the integral estimate and the error estimate associated with that region. The region is subdivided if the quadrature rule estimates for the integral has not met the required accuracy. The subdivided regions repeat the above steps recursively until the error estimate of the associated integration region meets the required accuracy. Many different adaptive integration methods have been developed in the past [5–9, 14]. Classical methods for 1-D adaptive integration include Simpson's method, Newton-Cotes 8-point method and Gauss-Kronrod 7/15-point and 10/21-point methods. Some of them have been extended to higher dimension [6].

An extension of 1-D quadrature rules for multidimensional integral is characterized by the exponential growth of functional evaluations with increasing dimension of integration region. For example, applying a Gauss-Kronrod 7/15-point

along each coordinate axis of a n-D integral requires 15^n evaluations of the integrand. Thus, an efficient integration algorithm for use in higher dimensions should be adaptive in the entire $n-$D space. CUHRE is one such open source algorithm which is available as a part of CUBA library [11, 12]. Even though the CUHRE method uses much fewer points, in practice it compares fairly well with other adaptive integration methods in terms of accuracy [10].

2.3 Overview of CUHRE

In this section we describe the sequential CUHRE algorithm for multidimensional integration. The integrals have the form

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2}\ldots\int_{a_n}^{b_n} f(\mathbf{x})d\mathbf{x}, \tag{6}$$

where $\mathbf{x}$ is an n-vector, and f is a scalar integrand. We use $[\mathbf{a},\mathbf{b}]$ to denote the hyper rectangle $[a_1,b_1]\times[a_2,b_2]\ldots\times[a_n,b_n]$.

The heart of the CUHRE algorithm is the procedure C-RULES$([\mathbf{a},\mathbf{b}],f,n)$ which outputs a triple (I,ε,κ) where I is an estimate of the integral over $[\mathbf{a},\mathbf{b}]$ (Equation 6), ε is an error estimate for I, and κ is the axis along which $[\mathbf{a},\mathbf{b}]$ should be split if needed. An important feature of C-RULES is that it evaluates the integrand only for $2^n+p(n)$ points where $p(n)$ is $\Theta(n^3)$ [5]. This is much fewer than 15^n function evaluations required by a straightforward adaptive integration scheme based on $7/15$-point Gauss-Kronrod method.

We now give a high-level description of the CUHRE algorithm (Algorithm 1). The algorithm input is n, $\mathbf{a}$, $\mathbf{b}$, f, a relative error tolerance parameter τ_{rel} and an absolute error tolerance parameter τ_{abs}, where $\mathbf{a}=(a_1,a_2,...,a_n)$ and $\mathbf{b}=(b_1,b_2,...,b_n)$. In the description provided below, H is a priority queue of 4-tuples $([\mathbf{x},\mathbf{y}],I,\varepsilon,\kappa)$ where $[\mathbf{x},\mathbf{y}]$ is a subregion, I is an estimate of the integral over this region, ε an estimate of the error and κ the dimension along which the subregion should be split if needed. The parameter ε determines the priority for extraction of elements from the priority queue. The algorithm maintains a global error estimate ε^g and a global integral estimate I^g. The algorithm repeatedly splits the region with greatest local error estimate and updates ε^g and I^g. The algorithm terminates when the $\varepsilon^g\leq\max(\tau_{abs},\tau_{rel}|I^g|)$ and outputs integral estimate I^g and error estimate ε^g.

Algorithm 1 SEQUENTIALCUHRE$(n,\mathbf{a},\mathbf{b},f,\tau_{rel},\tau_{abs})$

1: $(I^g,\varepsilon^g,\kappa)\leftarrow$ C-RULES$([\mathbf{a},\mathbf{b}],f,n)$
2: $H\leftarrow\emptyset$
3: INSERT$(H,([\mathbf{a},\mathbf{b}],I^g,\varepsilon^g,\kappa))$
4: **while** $\varepsilon^g>\max(\tau_{abs},\tau_{rel}|I^g|)$ **do**
5: $([\mathbf{a},\mathbf{b}],I,\varepsilon,\kappa)\leftarrow$ EXTRACT-MAX(H)
6: $a'\leftarrow(a_1,a_2,...,(a_\kappa+b_\kappa)/2,...,a_n)$
7: $b'\leftarrow(b_1,b_2,...,(a_\kappa+b_\kappa)/2,...,b_n)$
8: $(I_{left},\varepsilon_{left},\kappa_{left})\leftarrow$ C- RULES$([\mathbf{a},\mathbf{b}'],f,n)$
9: $(I_{right},\varepsilon_{right},\kappa_{right})\leftarrow$ C-RULES$([\mathbf{a}',\mathbf{b}],f,n)$
10: $I^g\leftarrow I^g-I+I_{left}+I_{right}$
11: $\varepsilon^g\leftarrow\varepsilon^g-\varepsilon+\varepsilon_{left}+\varepsilon_{right}$
12: INSERT$(H,([\mathbf{a},\mathbf{b}'],I_{left},\varepsilon_{left},\kappa_{left}))$
13: INSERT$(H,([\mathbf{a}',\mathbf{b}],I_{right},\varepsilon_{right},\kappa_{right}))$
14: **end while**
15: **return** $I^g$and ε^g

2.4 CUDA and GPU Architecture

Compute Unified Device Architecture (CUDA) [21] is a parallel computing platform and programming model for designing computations on the GPU. At the hardware level, a CUDA-enabled GPU device is a set of Single Instruction Multiple Data (SIMD) stream multi-processors (SM) with several stream processors (SP) each. Each SP has a limited number of registers and a private local memory. Each SM contains a global/device memory shared among the SPs within the same SM. Thread synchronization through shared memory is only supported between threads running on the same SM. Shared memory is managed explicitly by the programmers. The access to shared memory and register is much faster than access to global memory. The latency of accessing global memory is hundreds of clock cycles. Therefore, handling memory is an important optimization paradigm to exploit the parallel power of the GPU for general-purpose computing [15]. Besides the per-block shared memory and global memory, GPU device offers three other types of memory: per-thread private local memory, texture memory and constant memory. Texture memory and constant memory can be regarded as fast read-only caches.

CUDA programming model is a collection of *threads* running in parallel. A set of threads are organized as *thread blocks* and then, blocks are organized into *grids*. A grid issued by the host computer to GPU is called *kernel*. The maximum number of threads per block and number of blocks per grid are hardware-dependent. CUDA computation is often used to implement data parallel algorithms where for a given thread, its index is often used to determine the portion of data to be processed. Threads in common block communicate through shared memory. CUDA consists of a set of C language extensions and a runtime library that provides API to control the GPU device.

3 PARALLEL ADAPTIVE INTEGRATION METHODS

The sequential adaptive quadrature routine is poorly suited for GPUs. We propose a parallel algorithm that can utilize the

parallel processors of GPU to speed up the computation. The parallel algorithm approximates the integral by adaptively locating the subregions in parallel where the error estimate is greater than some user-specified error tolerance. It then calculates the integral and error estimates on these subregions in parallel. The pseudocode for the algorithm is provided below in the algorithms FIRSTPHASE (Algorithm 2) and SECONDPHASE (Algorithm 3).

Algorithm 2 FIRSTPHASE (n, $\mathbf{a}$, $\mathbf{b}$, f, d, τ_{rel}, τ_{abs}, L_{max})

1: $I^p \leftarrow 0$, $I^g \leftarrow 0$, $\varepsilon^p \leftarrow 0$, $\varepsilon^g \leftarrow \infty$
▷ I^p, ε^p - sum of integral and error estimates for the "good" subregions
▷ I^g, ε^g - sum of integral and error estimates for all subregions
2: $L \leftarrow$ INIT-PARTITION($\mathbf{a}, \mathbf{b}, L_{max}, n$)
3: **while** $(|L| < L_{max})$ and $(|L| \neq 0)$ and $(\varepsilon^g > \max(\tau_{abs}, \tau_{rel}|I^g|)$ **do**
4: $S \leftarrow \emptyset$
5: **for** all j in parallel **do**
6: $(I_j, \varepsilon_j, \kappa_j) \leftarrow$ C-RULES($L[j], f, n$)
7: INSERT($S, (L[j], I_j, \varepsilon_j, \kappa_j)$)
8: **end for**
9: $L \leftarrow$ PARTITION($S, L_{max}, \tau_{rel}, \tau_{abs}$)
10: $(I^p, \varepsilon^p, I^g, \varepsilon^g) \leftarrow$ UPDATE($S, \tau_{rel}, \tau_{abs}, I^p, \varepsilon^p$)
11: **end while**
12: **return** ($L, I^p, \varepsilon^p, I^g, \varepsilon^g$)

Algorithm 3 SECONDPHASE(n, $\mathbf{f}$, τ_{rel}, τ_{abs}, L, I^g, ε^g)

1: **for** $j = 1$ to $|L|$ **parallel do**
2: Let $[\mathbf{a_j}, \mathbf{b_j}]$ be the j^{th} record in L
3: $(I_j, \varepsilon_j) \leftarrow$SEQUENTIALCUHRE($n$, $\mathbf{a_j}$, $\mathbf{b_j}$, f, τ_{rel}, τ_{abs})
4: **end for**
5: $I^g \leftarrow I^g + \sum_{[\mathbf{a_j}, \mathbf{b_j}] \in L} I_j$
6: $\varepsilon^g \leftarrow \varepsilon^g + \sum_{[\mathbf{a_j}, \mathbf{b_j}] \in L} \varepsilon_j$
7: **return** $I^g$and ε^g

3.1 Adaptive Multidimensional Integration On a Single GPU

FIRSTPHASE: The goal of FIRSTPHASE is to create a list of subregions of the whole region $[\mathbf{a}, \mathbf{b}]$, with at least L_{max} elements for which further computation is necessary for estimating the integral to desired accuracy (L_{max} is a parameter that is based on target GPU architecture). This list is later passed on to SECONDPHASE. The algorithm maintains an list L of subregions, stored as $[\mathbf{a_j}, \mathbf{b_j}]$. Initially the whole integration region is split into roughly L_{max} equal parts through the procedure INIT-PARTITION. In each iteration of the while loop in FIRSTPHASE, first the CUHRE rules are applied to all subregions in L in parallel to get the integral estimate, error estimate, and the split axis. A list S is created to store the intervals with these values. Thereafter the algorithm essentially identifies the "good" and the "bad" subregions in S – the good subregions have error estimate that is below a chosen threshold, whereas bad subregions have error estimates exceeding this threshold. The bad subregions need to be further divided, while the integral and error estimates for the good regions can simply be accumulated. This is accomplished through the procedures PARTITION and UPDATE. Pseudocode for these procedures are provided in [2].

First phase continues until (i) a long enough list of "bad" subregions is created in which case we proceed to the second phase or (ii) there are no more "bad" subregions in which case we can return the integral and error estimates I^g and ε^g as the answer or (iii) I^g, ε^g satisfy the error threshold criteria in which case we also return I^g and ε^g as the answer. Note that, in case (ii) or (iii) second phase of the algorithm is not used.

In our GPU implementation, FIRSTPHASE kernel implements the C-RULE on every GPU thread to estimate the triplet (I, ε, κ) for a subregion assigned to it. This kernel requires at least as many threads as there are subregions in the input list and creating multiple threads hide the latency of global memory by overlapping the execution. The kernel returns a list of triplets computed by each thread along with a identifier which specifies if a subregion has to be further subdivided or not. The intermediate integral estimates are evaluated as the sum of individual estimates for all subregions in the list. We make use of CUDA-based THRUST library [3, 18] to perform such common numerical operations. All the bad subregions are identified and copied to a new list based on the identifier flag. Prefix scan [20] implementation from the CUDA THRUST library is used to identify the position of bad subregions in the subregion list. Identified bad subregions are further partitioned into finer subregions, and the implementation continues with the steps above on these finer subregions. Details of this GPU implementation is described in [2].

SECONDPHASE: In SECONDPHASE, on every subregion $[\mathbf{a_j}, \mathbf{b_j}]$ in the list L the algorithm calls sequential CUHRE routine (SEQUENTIALCUHRE) to compute global integral and error estimate for the selected subregion. SECONDPHASE implements the modified version of SEQUENTIALCUHRE on every GPU block to estimate the integral value for the subregion assigned to it. This means that *for* loop in Algorithm 3 is parallelized such that each subregion is mapped to a block of threads. A block in the SECONDPHASE kernel works independent of the other blocks in estimating the integral value of the subregion assigned to them. Each block maintains a list of subregion record in the memory, which is split between per-block shared memory and global memory. The

shared memory here functions as a cache to global memory in storing a partial list of subregion records with higher error estimates. The C-RULE function evaluations in the SEQUENTIALCUHRE procedure are distributed equally among the available threads in a block. Details of this GPU implementation is described in [2], [1].

We use the constant memory to store the C-RULE parameters that do not change during the algorithm execution such as evaluation points on a unit hypercube, and the corresponding weights. This provides a significant performance improvement as compared to storing them in global memory. Before invoking the GPU kernels on a set of subregions, all these C-RULE parameters are loaded into constant memory. The 64KB memory capability of the constant memory limits the number of parameters that can be stored in the constant space. Structure and representation of C-RULE parameters are optimized to best fit in the available constant memory.

3.2 Adaptive Multidimensional Integration On a Multiple GPUs

The general idea is to extend the single GPU algorithm across a cluster of compute nodes with multiple GPU devices per node. This involves dividing the subregions generated by the FIRSTPHASE kernel equally among the available GPU devices and implementing the SECONDPHASE kernel on each of these device. The FIRSTPHASE algorithm here creates a list of subregions for the whole region $[\mathbf{a}, \mathbf{b}]$, with at least L_{max} elements for which further computation is necessary for estimating the integral to desired accuracy. The optimal value of L_{max} is estimated based on the target architecture and the number of available GPU devices. For our implementation we have used $L_{max} = 2048d$, where d is the number GPU devices. The generated list of subregions are equally partitioned among the available GPU devices and each partition is assigned to a GPU device implementing the SECONDPHASE kernel.

Communication between GPU devices attached to a compute node are handled using OpenMP, whereas the communication between the compute nodes are handled using MPI programming. All the memory transfers between GPU devices at a node are done using the host (compute node) as an intermediary. The algorithm starts by creating a MPI process for each compute node. One of the MPI process initializes the C-RULE parameters and implements the FIRSTPHASE on a single GPU device to generate a list of subregions. The generated list is transferred to the host memory where it is partitioned equally among the available computes nodes. Each of these partitions are distributed to the compute nodes using MPI routines. Compute nodes in the cluster receives a set of subregions from the node implementing FIRSTPHASE. These subregions are further partitioned among the available GPU devices at the compute node. Using OpenMP routines, each node creates a thread for every GPU device attached to it. A thread running at the compute node initializes the assigned GPU device and transfers the subregion list to the GPU device memory. SECONDPHASE is executed is parallel by all the threads at the compute node. After completion of SECONDPHASE, the results are transferred back to the node implementing FIRSTPHASE using MPI routines. In our implementation we make use of CUDA-based THRUST library [3, 18] to perform common numerical operations such as summation and prefix scan [20]. These operations are often used in reducing the results obtained from each device.

The scalability of multi-GPU approach often depends on the cluster size and nature of integrand. When the integrand converges to the required accuracy during the FIRSTPHASE, the SECONDPHASE is never used. The overall performance for such integrands is not affected by the cluster size beyond a threshold. However, this scenario is not common with the poorly behaved integrands that is often encountered in science.

4 PERFORMANCE/EXPERIMENTAL RESULTS

Our experiments for single GPU versions were carried out on a NVIDIA Tesla M2090 GPU device installed on a compute node (host) with Intel® Xeon® CPU X5650, 2.67GHz. A Tesla M2090 offers 6GB of GDDR5 on-board memory and 512 streaming processor cores (1.3 GHz) that delivers a peak performance of 665 Gigaflops in double precision floating point arithmetic. The interconnection between the host and the device is via a PCI-Express Gen2 interface. The experiments for multiple GPU approach were carried out on a cluster of 6 Intel® Xeon® CPU X5650 computes nodes with 4 NVIDIA Tesla M2090 GPU devices on each compute node. The GPU code was implemented using CUDA 4.0 programming environment. The source code of our multi-GPU implementation is made available at `https://github.com/akkamesh/GPUComputing`.

We carried out our evaluation on a set of challenging functions which require many integrand evaluations for attaining the prescribed accuracy. We use the battery of benchmark functions (Table 1) which is representative of the type of integration that is often encountered in science: oscillatory, strongly peaked and of varying scales. These kinds of poorly-behaved integrands are computationally costly, which is why they greatly benefit from a parallel implementation. The region of integration for all the benchmark functions in our experiments is a unit hypercube $[0, 1]^n$. For comparison, we use the sequential C-implementation of CUHRE from the CUBA package [11, 12] that was executed on the host machine.

Table 2 compares the performance results for sequential implementation, GPU implementation with one and 24 devices for a subset of functions from the benchmark suite. The

Function	n	τ_{rel}	Sequential Time (sec.)	Single GPU Time (sec.)	Single GPU Speedup	24 GPUs Time (sec.)	24 GPUs Speedup
$f_1(\mathbf{x})$	7	10^{-5}	2349	18.12	129.63	3.87	607.18
$f_2(\mathbf{x})$	4	10^{-4}	3621	15.10	239.92	1.11	3252.6
$f_3(\mathbf{x})$	4	10^{-5}	286	11.36	25.19	4.55	62.89
$f_4(\mathbf{x})$	7	10^{-4}	9876	71.75	137.65	19.60	503.90

Table 2: Performance results for sequential implementation on CPU, GPU implementation with one and 24 GPU devices for benchmark functions in Table 1.

1. $f_1(\mathbf{x}) = \left[\alpha + \cos^2\left(\sum_{i=1}^{n} x_i^2\right)\right]^{-2}$, where $\alpha = 0.1$

2. $f_2(\mathbf{x}) = \cos\left(\prod_{i=1}^{n} \cos\left(2^{2i} x_i\right)\right)$

3. $f_3(\mathbf{x}) = \sin\left(\prod_{i=1}^{n} i \arcsin(x_i^i)\right)$

4. $f_4(\mathbf{x}) = \sin\left(\prod_{i=1}^{n} \arcsin(x_i)\right)$

Table 1: n-D benchmark functions.

dimension and accuracy for these functions are chosen based on the highest value of these parameters at which the sequential implementation was able to compute the results before reaching the limit for total function evaluations of 10^9. The speedup here is computed by comparing the total execution time for the parallel code on GPU against the time taken by serial code on the compute node. We observe that the GPU implementation is up to 240 times faster than the serial code on a single GPU device.

In Figure 1 we show the impact on the speedup with the number of GPU devices for two different functions: $f_2(\mathbf{x})$ in 4-D space with a relative error of $\tau = 10^{-4}$ and $f_1(\mathbf{x})$ in 7-D space with relative error of $\tau = 10^{-4}$. These two functions are chosen to illustrate different behaviors of all the simulations executed. The speedup here is with reference to the implementation on single GPU device. We observe a speedup of up to 14 on 24 GPU devices compared against one GPU device. This translates to a overall speedup of up to 3250 with 24 GPUs compared to a sequential implementation.

With the increase in number of GPUs, FIRSTPHASE kernel generates more balanced computational load and thus improving the performance. In Figure 2, we show the effectiveness of having two phases in our algorithm by comparing the results of the implementation with FIRSTPHASE against the one without FIRSTPHASE. Figure 2a and Figure 2c show the result of executing two-phase GPU algorithm without FIRSTPHASE and Figure 2b and Figure 2d show the normal execution with FIRSTPHASE. Both these evaluations were performed on a 5-D function $f_3(\mathbf{x})$ chosen from the benchmark with a relative error requirement of 10^{-2} and 10^{-3}.

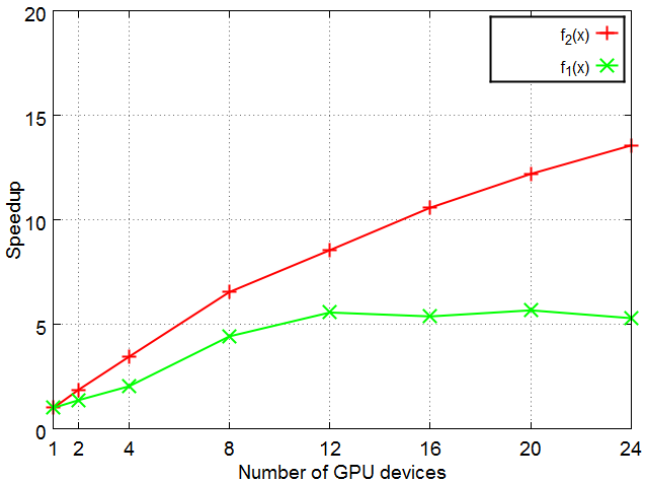
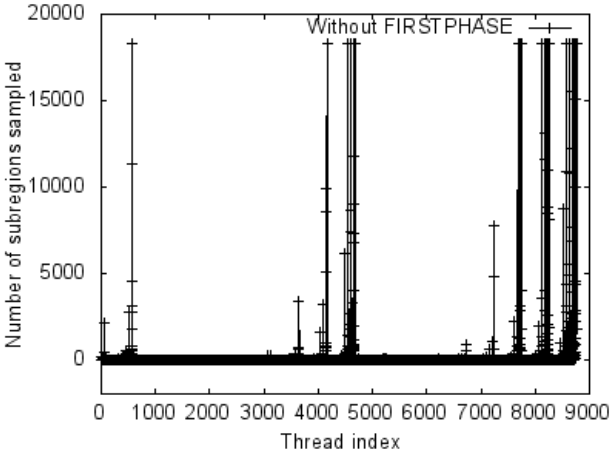
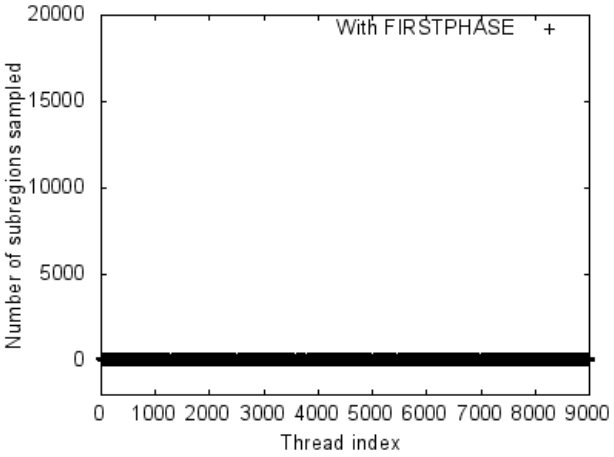


Figure 1: Speedup results for the GPU implementation with varying number of GPUs for two functions: $f_1(\mathbf{x})$ in 7-D and $f_2(\mathbf{x})$ in 4-D (speedup is with reference to the implementation on single GPU).

In each of these figures we plot the number of subregions sampled by a thread in SECONDPHASE against the thread index. Computational load of a thread here is directly related to the number of subregions sampled by that thread. GPUs that are built on SIMD architecture require every thread to share approximately equal load to gain maximum performance. In Figure 2a and Figure 2c, we observe a wide variance of subregions sampled by the threads. Some of these threads have longer execution time than others, which results in an unbalanced computational load. The overall execution time greatly depends on these threads which have longer execution times. This brings out the importance of FIRSTPHASE to share the load across the threads. Figure 2b and Figure 2d show the execution of SECONDPHASE with the FIRSTPHASE behaving as a load balancer. We notice that the number of subregions sampled by the threads are approximately same, reflecting a efficient load balancing. The total execution time in both cases – with or without the FIRSTPHASE – depends on the execu-

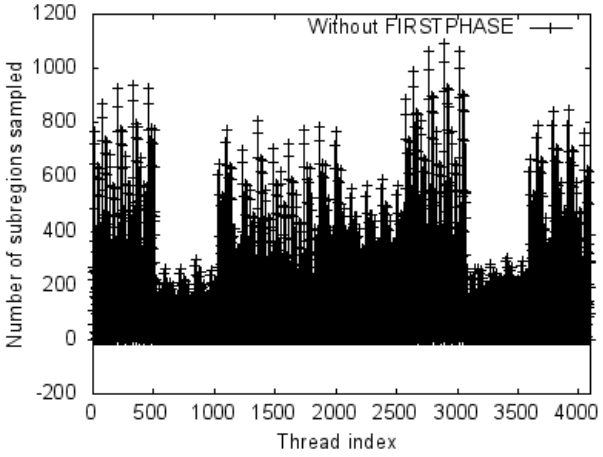


(a) Without FIRSTPHASE for $f_3(\mathbf{x})$ with $\tau_{rel} = 10^{-2}$ and $n = 5$.

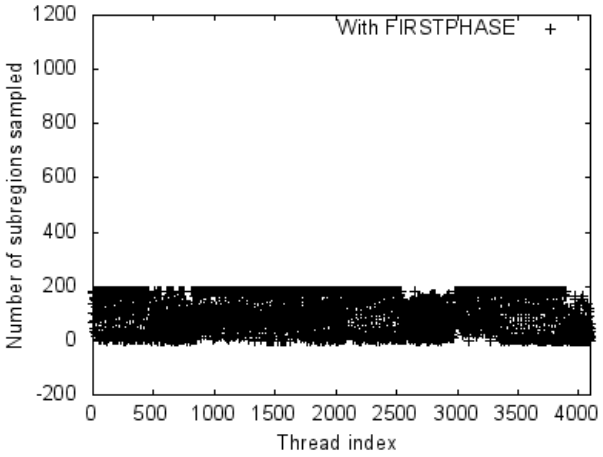


(b) With FIRSTPHASE for $f_3(\mathbf{x})$ with $\tau_{rel} = 10^{-2}$ and $n = 5$.

Figure 2: GPU results for execution with FIRSTPHASE and without FIRSTPHASE.



(c) Without FIRSTPHASE for $f_3(\mathbf{x})$ with $\tau_{rel} = 10^{-3}$ and $n = 5$.



(d) With FIRSTPHASE for $f_3(\mathbf{x})$ with $\tau_{rel} = 10^{-3}$ and $n = 5$.

Figure 2: GPU results for execution with FIRSTPHASE and without FIRSTPHASE.

tion time of the most highly loaded thread, which in the case when FIRSTPHASE serves as a load balancer is considerably shorter (Figure 2a and Figure 2c). We notice that due to the nature of GPUs, we obtain higher performance by having two phases.

5 DISCUSSION AND CONCLUSION

We presented a two-phase algorithm for solving multidimensional numerical integration on a cluster of compute nodes with multiple GPU devices per node. We obtained a speedup of up to 240 on a single GPU device and on a cluster of 6 nodes (24 GPU devices) we were able to obtain a speedup of up to 3250 compared to a leading sequential method. The presented algorithm is a generic solution for multidimensional numerical integration which can be widely used in various fields of computational science, such as lattice QCD simulations, simulation of CSR in charged particle beam, solution of the Navier-Stokes equations using spectral element methods requiring the ability to perform multidimensional integration for billions of points, quantum mechanics calculations and others.

Our current and future efforts are focused on extending the two-phase algorithm to fit in a CSR model and furthermore develop a highly efficient simulation model for CSR using state-of-the art computing platforms consisting of CPUs and GPUs. Using NVIDIA's CUDA framework, several of the new CSR model's vital numerical algorithms, such as multidimensional interpolation and charge deposition, will be designed to run optimally on such hybrid platforms. Our new simulation model will not only defray the operation cost but also contribute to fine-tuning the parameters for more efficient operation of electron accelerators.

ACKNOWLEDGMENT

We would like to thank the support of the Jefferson Science Associates Project.712336 and the U.S. Department of Energy (DOE) Contract No. DE-AC05-06OR23177.

REFERENCES

[1] K. Arumugam, A. Godunov, D. Ranjan, B. Terzić, and M. Zubair. A Memory-Efficient Algorithm for Adaptive Multidimensional Integration with Multiple GPUs. *International Conference on High Performance Computing (HiPC)*, December 2013.

[2] K. Arumugam, A. Godunov, D. Ranjan, B. Terzić, and M. Zubair. An Efficient Deterministic Parallel Algorithm for Adaptive Multidimensional Numerical Integration on GPUs. *International Conference on Parallel Processing (ICPP)*, October 2013.

[3] N. Bell and J. Hoberock. Thrust library for GPUs.

[4] J. Bernsten. Adaptive-multidimensional quadrature routines on shared memory parallel computers. *Reports in Informatics 29, Dept. of Informatics, Univ. of Bergen*, 1987.

[5] J. Bernsten, T. Espelid, and A. Genz. An adaptive algorithm for the approximate calculation of multiple integrals. *ACM Transactions on Mathematical Software (TOMS)*, 17(4):437–451, December 1991.

[6] Germund Dalquist and Åke Björck. *Numerical Methods in Scientific Computing*, volume 1. Society for Industrial and Applied Mathematics, 2008.

[7] Alan Genz. An adaptive multidimensional quadrature procedure. *Computer Physics Communications*, 4:11–15, October 1972.

[8] Alan Genz and Ronald Cools. An adaptive numerical cubature algorithm for simplices. *ACM Transactions on Mathematical Software (TOMS)*, 29(3):297–308, September 2003.

[9] Alan Genz and A.A. Malik. An adaptive algorithm for numerical integration over an n-dimensional rectangular region. *Journal of Computational and Applied Mathematics*, 6:295–302, December 1980.

[10] Pedro Gonnet. A review of error estimation in adaptive quadrature. *ACM Computing Surveys (CSUR)*, 44(22), December 2012.

[11] T. Hahn. CUBA - The CUBA library. *Nuclear Instruments and Methods in Physics Research*, 559:273–277, 2006.

[12] T. Hahn. CUBA - a library for multidimensional numerical integration. *Computer Physics Communications*, 176:712–713, June 2007.

[13] IMSL. International mathematical and statistical libraries. Rogue Wave Software, 2009.

[14] J. Bernsten and T. Espelid and A. Genz. DCUHRE: an adaptive multidemensional integration routine for a vector of integrals. *ACM Transactions on Mathematical Software (TOMS)*, 17(4):452–456, December 1991.

[15] Yooseong Kim and Aviral Shrivastava. CuMAPz: A tool to analyze memory access patterns in CUDA. *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, pages 128 – 133, June 2011.

[16] R. Li. Self-consistent simulation of the CSR effect on beam emittance. *Nuclear Instruments and Methods in Physics Research Section A*, 429:310–314, 1998.

[17] R. Li. The Physics of High Brightness Beams . Proceedings of the 2nd ICFA Advanced Accelerator Workshop, 1999.

[18] N. Bell and J. Hoberock. Thrust: A productivity-oriented library for CUDA. *GPU Computing Gems Jade Edition*, 2011.

[19] NAG. Fortran 90 Library. Numerical Algorithms Group Inc., Oxford, U.K., 2000.

[20] Hubert Nguyen. Parallel prefix sum (scan) with CUDA. *GPU Gems 3*, 2007.

[21] NVIDIA. CUDA C Programming Guide.

[22] R. Piessens and E. de Doncker-Kapenga and C. Überhuber, and D. Kahaner. *QUADPACK: A Subroutine Package for Automatic Integration*. Springer-Verlag, Berlin, 1983.

Modeling Protein Structure Features from Three Dimensional Cryo-EM Images

Dong Si, Jing He
Computer Science, Old Dominion University
dsi@cs.odu.edu, jhe@cs.odu.edu

Keywords: Cryo-electron microscopy, three-dimensional density image, feature recognition, computer visualization, protein, beta-sheet structure, surface modeling

Abstract

Secondary structure of protein, such as α-helix and β-sheet, is the general three-dimensional (3D) form of local segments. It can be identified from the 3D electron cryo-microscopy (cryo-EM) density images at medium resolutions (~5-10Å). A detected β-sheet can be represented by either the voxels of β-sheet density or by many piecewise polygons to compose a rough surface. However, none of these is effective in capturing the global surface feature of the β-sheet. In addition to the single layer sheet, β-barrel as a particular sheet structural feature is formed by multiple β-strands in a barrel shape. We present a novel mathematical model to represent the single layer β-sheet density, and also an optimized model to represent the β-barrel density. These surface models can be potentially used for further detection of β-strands when the resolution is not high enough to resolve the molecular details, it is critical for the de-novo backbone structure derivation in cryo-EM density images at the medium resolutions.

1. INTRODUCTION

Electron cryo-microscopy (Cryo-EM) has become a major experimental technique to study the structures of large protein complexes, such as ribosomes and viruses [1, 2]. It is a structure determination technique complementary to the X-ray Crystallography and Nuclear Magnetic Resonance (NMR). At the medium resolutions such as 5-10Å, detailed molecular features are not resolved. However, secondary structure features such as α-helices and β-sheets can be computational identified (Figure 1). The α-helix appears as a stick (red in Figure 1) and can be identified using image processing methods [3-6]. A β-sheet appears as a thin layer of density (blue in Figure 1) and can be detected computationally [5-8]. Some β-sheets curve into β-barrels (Figure 1E and F). A β-barrel is composed of multiple β-strands (ribbon of Figure 1F) that twist and coil to form a closed structure in which the first strand is hydrogen bonded to the last.

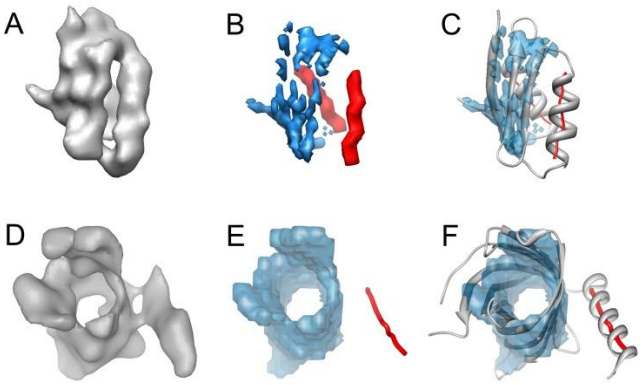


Figure 1. The secondary structure detection from 3-dimensinal protein density images, Protein denisty image simulated were using EMAN [9]. (A) The density image of protein 2AW0 at 8Å resolution; (B) the helix (red) and β-sheet (blue) voxels detected by *SSELearner* [6]; (C) the computationally detected secondary structures superimposed on the PDB structure (shown in ribbon); (D) the density image of protein 3GP6 at 10Å resolution; (E) the detected helix (red line) and β-barrel region (blue voxels) using *SSETracer* [8]; (F) the atomic structure of the protein (shown in ribbon) overlaped with the detected secondary structures.

The accuracy of secondary structure identification from volumetric protein density images is critical for de-novo backbone structure derivation in cryo-EM. It is still challenging to detect the secondary structures automatically and accurately from the density images at medium resolutions (~5-10Å). We have developed a machine learning approach, *SSELearner*, to automatically identify α-helices and β-sheets by using a combination of image processing and supervised machine learning techniques [6].

Based on the local shape characteristics of α-helices and β-sheets, we have also developed *SSEtracer*, which performs a series of local feature analysis to detect the secondary structures [8]. A helix detected from the medium resolution data is often represented as a spline (red line in Figure 1C and F), referred as an α-trace that corresponds to the central axis of the helix. A detected β-sheet can be represented by either the β-sheet density voxels (Figure 1C and F) or by many piecewise polygons to compose a rough surface [5]. However, none of these is effective in capturing the global surface feature of the β-sheet. Without modeling the entire β-sheet as an accurate surface model, the detection of β-strands would be difficult.

2. METHODOLOGY

2.1. Single Layer Beta-sheet Surface Modeling

Many studies have shown that a variety of saddle shaped surfaces can be used to model β-sheets in atomic structures. Helicoids have been used to fit small β-sheets using the principle of minimal surfaces [10]. Additional models involve catenoid for β-sandwiches [11].

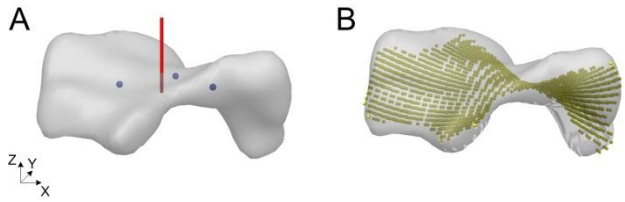


Figure 2. Fitted polynomial surface to the β-sheet density. (A) The center plane that decided by three cluster points (blue balls) with its normal vector (red line); (B) fitted 3D surface model (yellow surface points were generated by the model).

Rather than using different forms for different types of surface, a more general model was proposed for polynomial surface [12]. Although the order-two polynomial surface can already describe the surface pattern for β-sheets, we decided to use order-three 3D polynomial surface (Equation 3), in order to capture the flexibility and curvature for β-sheet density at medium resolutions.

$$z = Ax^3 + By^3 + Cx^2 + Dy^2 + Ex^2y + Fy^2x + Gxy + Hx + Iy + J \quad (1)$$

Since Equation (1) is a function that maps coordinate x and y to coordinate z, the 3-dimensional surface can be best fitted when the β-sheet density area is approximately parallel to $X - Y$ plane and the approximate normal vector of β-sheet density is roughly along the Z direction. Due to the folded shape of β-sheet, the geometry center of β-sheet density may not be on the density itself. We first searched some scattered cluster points from the density voxels based on a distance cutoff 5Å, and defined the sheet center as the closest voxel to the density geometry center. The three cluster points that are closest to the sheet center were picked to build a center plane for finding the rough normal vector of the β-sheet density (Figure 2A). The β-sheet density was then rotated so that the normal vector of sheet density is aligned with the Z direction (Figure 2A). The β-sheet density area was then fitted with the polynomial surface model (Equation 1) using least-square method, as shown in Figure 2B. The (x, y, z) in this equation is the coordinate of the β-sheet density voxel. All the ten coefficients in this equation can be optimized using least-square fitting method. Finally, the β-sheet density was rotated back after the modeling has done.

2.2. Beta-barrel Surface Modeling

β-barrels have characteristic shapes and have been modeled mathematically in previous studies. The atomic structure of a β-barrel has been modeled as hyperboloid surfaces [13-15] and catenoid surfaces [16]. All these methods concentrated on the fitting of a particular mathematical model to the β-barrel structures by using linear or non-linear fitting procedure. Although these models approximate the majority area of a β-barrel, the images of β-barrels often deviate from the rigid mathematical models in certain area. We present an adaptive method to generate the surface that fit in the 3-dimensional image of a β-barrel density. The idea is to use a rigid model for area that fit well and then optimize the model on where it does not fit.

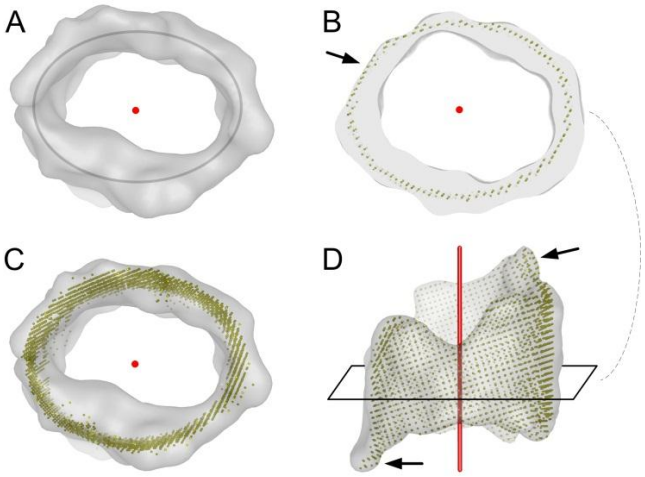


Figure 3. Modeling the surface from 3D β-barrel density image. (A) The barrel axis (red) was searched by fitting an elliptical cylinder (line) to the density image (gray), shown as the top view; (B) one cross-section of the barrel, arrow shows the shrinking area of the surface model according to the morphed density; (C) top view of the modeled β-barrel surface (yellow); (D) side view of the barrel surface that modeled from the density and the axis (red).

For the region of density that related to a β-barrel, least-square procedure was first performed to find the central axis of the barrel by fitting an elliptical cylinder to it (Figure 3A).

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (2)$$

The purpose of fitting a cylinder here is to find the Z axis of the β-barrel. Then we built the surface model of the β-barrel from bottom to top. The density voxels on each cross-section of Z axis (Figure 3B, gray) look like a round belt, and the cross-section of the fitted elliptical cylinder on each Z axis is an ideal ellipse. We saved the voxels that roughly around the ellipse (Figure 3B, yellow) to the surface model, if the ellipse is within the density voxels on each cross-

section. For the region where the fitted elliptical cylinder is outside the density (arrow in Figure 3B), we searched the closest voxels to the ellipse and saved those voxels to the surface model. The surface model was built one layer by one layer until reach the top of the β-barrel density. Our modeled barrel surface clearly follows to the morphed regions (arrows in Figure 3D) of the density image.

3. RESULTS

We tested our single-sheet surface modeling method on the β-sheet density that detected from experimental derived cryo-EM density images by *SSEtracer*. The experimental derived cryo-EM density images often contain noises and bump at the edge area of β-sheet because of closeness to parts of other structures such as loops or turns (Figure 4). Our polynomial model of β-sheet represents the overall twisted surface pattern (Figure 4, right column). The surface points shown in Figure 4 are generated by the polynomial model (Equation 1). As an example, the ten coefficients that calculated for the polynomial surface model of β-sheet shown in Figure 4A are listed: $A = 0.0013, B = 0.0022, C = 0.0017, D = -0.0004, E = 0.0731, F = -0.0084, G = 0.0699, H = 3.2652, I = -1.8834, J = 1.9414$. In this model, each parameter ($A \ldots J$) can be associated with a feature of the 3D surface. For example, the combinations of $A \ldots G$ produce more complex of surfaces while H and I simply tilt the surface in x and y respectively, and J sets the base level. As shown in Figure 4, the polynomial surface model visually fits in the detected β-sheet cryo-EM density area well and represents the 3D surface feature of the β-sheets.

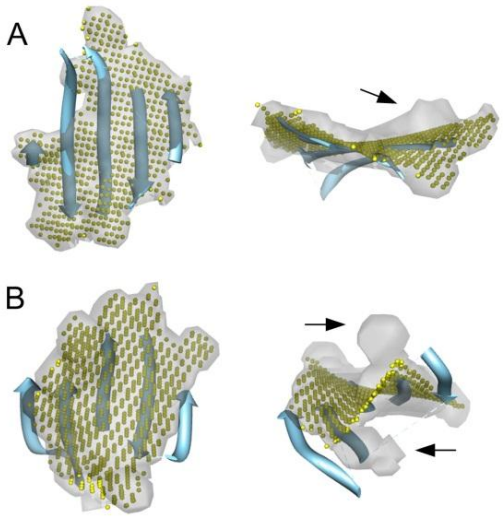


Figure 4. Polynomial model that fits in the β-sheet density. (A) Generated points (yellow) by the polynomial model to show the 3D surface, superimposed on the true structure (cyan ribbon, sheet A of protein 3C92) and the detected sheet density (gray, EMDB entry 1740); (B) sheet W of protein 3IZ6 and the detected sheet density from EMDB entry 1780.

We also tested our β-barrel surface modeling method on the β-barrel density images. As shown in Figure 5, the optimized surface model fits in the β-barrel density voxels well and represents the 3D feature of the β-barrel as a thin surface. Our modeled surface of β-sheet and β-barrel is a simplified representation over the density voxel representation and other piecewise polygon representation. More detail of the geometry features that hidden inside the chunk of density area can be then revealed by using these models.

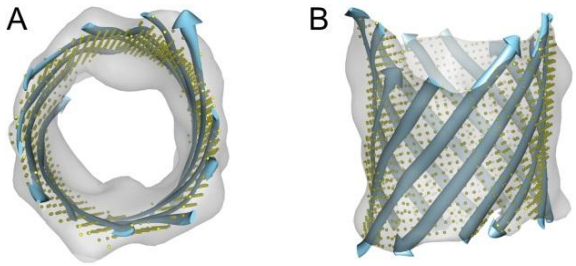


Figure 5. Suface model that built from the β-barrel density. (A) Generated points (yellow) to show the 3D surface in top view, superimposed on the true structure (cyan ribbon, sheet A of protein 4FQE) and the β-barrel density (gray); (B) is the side view of (A).

One of the significant contributions of the β-sheet surface model is for identifying the β-strands. This is due to the simplicity of the model yet capturing the overall curvature of the β-sheet. Figure 6 shows an example of the β-strands (red curve) that calculated from our β-sheet and β-barrel surface models. This top ranked detection of β-strand aligns well with the true β-strands (blue ribbon). The details of our β-strand detection method are included in the separated papers.

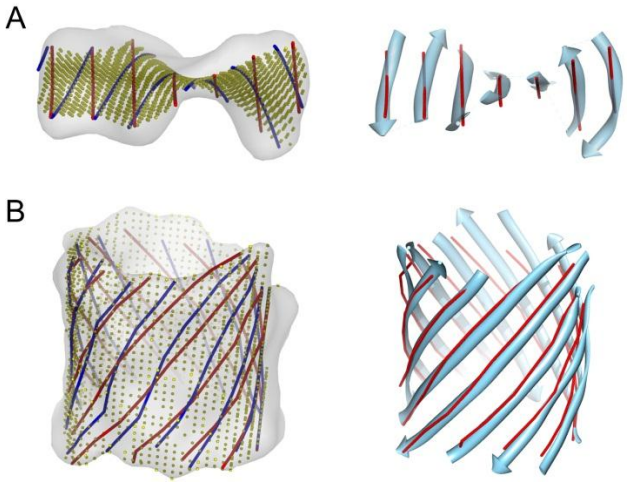


Figure 6. Detected β-strands using the surface model of β-sheet and β-barrel. (A) 3D polynomial β-sheet surface model (yellow) and the possible β-strand samples (blue and red curves) that on the modeled surface, the best detected β-strand position (red curve) is superimposed on the true structure

(cyan ribbon); (B) the β-barrel surface model and the detected β-strand positions, shown as the same color scheme in (A).

4. CUNCLUSION

We proposed a polynomial model for the single layer β-sheet density and an adaptive surface model for the β-barrel density. The models can be used to represent the overall surface feature of β-sheets and β-barrels, and it is a simplified representation over the voxel representation and other piecewise polygon representation. We gave an example to show how these models could be further assisted in the detection of β-strand location and orientation, which is critical for the de-novo backbone structure derivation in cryo-EM density images at the medium resolutions.

Biography

Dong Si received his B.S. in Electronic Information Science and Technology with honors from Nanjing University, China. In fall 2009, Dong joined the Computer Science Department of Old Dominion University as a graduate student and then transferred to Ph.D. program in fall 2010. He is currently working with Dr. Jing He as a Research Assistant. Dong has been offered the "Modeling and Simulation Research Fellowship" from Virginia Modeling, Analysis and Simulation Center (VMASC) in 2013.

Jing He obtained her B.S. degree in Mathematics from Jilin University and M.S. degree in Mathematics from New Mexico State University. She worked in the area of three-dimensional reconstruction and analysis of virus structures at Baylor College of Medicine from which she obtained her Ph.D. in Structural and Computational Biology and Molecular Biophysics. Currently she is an Associate Professor at the Department of Computer Science at Old Dominion University.

References

[1] W. Chiu, M. L. Baker, W. Jiang, M. Dougherty, and M. F. Schmid, "Electron cryomicroscopy of biological machines at subnanometer resolution," *Structure*, vol. 13, pp. 363-72, Mar 2005.

[2] Z. H. Zhou, "Atomic resolution cryo electron microscopy of macromolecular complexes," *Adv Protein Chem Struct Biol*, vol. 82, pp. 1-35, 2011.

[3] W. Jiang, M. L. Baker, S. J. Ludtke, and W. Chiu, "Bridging the information gap: computational tools for intermediate resolution structure interpretation," *J Mol Biol*, vol. 308, pp. 1033-44, May 2001.

[4] A. Del Palu, J. He, E. Pontelli, and Y. Lu, "Identification of Alpha-Helices from Low Resolution Protein Density Maps," presented at the Proceeding of Computational Systems Bioinformatics Conference(CSB), 2006.

[5] M. L. Baker, T. Ju, and W. Chiu, "Identification of secondary structure elements in intermediate-resolution density maps," *Structure*, vol. 15, pp. 7-19, Jan 2007.

[6] D. Si, S. Ji, K. A. Nasr, and J. He, "A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps," *Biopolymers*, vol. 97, pp. 698-708, Sep 2012.

[7] Y. Kong and J. Ma, "A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps," *J Mol Biol*, vol. 332, pp. 399-413, Sep 12 2003.

[8] D. Si and J. He, "Beta-sheet Detection and Representation from Medium Resolution Cryo-EM Density Maps," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ed. Wshington DC, USA: ACM, 2013, pp. 764-770.

[9] S. J. Ludtke, P. R. Baldwin, and W. Chiu, "EMAN: semiautomated software for high-resolution single-particle reconstructions," *J Struct Biol*, vol. 128, pp. 82-97, Dec 1 1999.

[10] D. Znamenskiy, K. Le Tuan, A. Poupon, J. Chomilier, and J. P. Mornon, "Beta-sheet modeling by helical surfaces," *Protein Eng*, vol. 13, pp. 407-12, Jun 2000.

[11] E. Koh and T. Kim, "Minimal surface as a model of beta-sheets," *Proteins-Structure Function and Bioinformatics*, vol. 61, pp. 559-569, Nov 15 2005.

[12] W. R. Taylor and A. Aszodi, *Protein geometry, classification, topology and symmetry : a computational analysis of structure*. Bristol: Institute of Physics Pub., 2005.

[13] J. Novotny, R. E. Bruccoleri, and J. Newell, "Twisted hyperboloid (Strophoid) as a model of beta-barrels in proteins," *J Mol Biol*, vol. 177, pp. 567-73, Aug 15 1984.

[14] E. Tolonen, B. Bueno, S. Kulshreshta, P. Cieplak, M. Argaez, L. Velazquez, and B. Stec, "Allosteric transition and binding of small molecule effectors causes curvature change in central beta-sheets of selected enzymes," *J Mol Model*, vol. 17, pp. 899-911, Apr 2011.

[15] I. Lasters, S. J. Wodak, P. Alard, and E. van Cutsem, "Structural principles of parallel beta-barrels in proteins," *Proc Natl Acad Sci U S A*, vol. 85, pp. 3338-42, May 1988.

[16] E. Koh and T. Kim, "Minimal surface as a model of beta-sheets," *Proteins*, vol. 61, pp. 559-69, Nov 15 2005.

LINKS BETWEEN OPERATIONS RESEARCH AND MODELING & SIMULATION

Daniele Vernon-Bido and Andrew Collins

ABSTRACT

The analysis portion of modeling and simulation shares many links with operations research. Both disciplines came into existence in the 1930s. Both are multidisciplinary and practical in nature. They share the challenges of model validity, usability and acceptability. As modeling and simulation specialists struggle to define the discipline, OR academics deal with the challenge to maintain relevance. This paper explores the history of both OR and M&S and their shared paradigm in an attempt to help shape the future.

KEYWORDS: operations research, modeling and simulation, constructivism

1. INTRODUCTION

Modeling and Simulation (M&S) as a tool and topic of research has existed for more than 70 years [1]. Yet the paradigms of the discipline (or even if M&S is a discipline) is still debated by researchers and practitioners. Operations Research (OR) is an established discipline that has been around just as long. Given that each discipline is both practical and technical, it is not surprising to find links between them.

M&S is often characterized by its objectives: system analysis, education and training, acquisition and system acceptance, research and entertainment [1]. System analysis attempts to model the behavior of a system to improve understanding or performance of the system. Education and training is used to aid the understanding and application of concepts. Acquisition and system acceptance is intended to provide information about system performance and subsystem integration. Research creates artificial environments with components that can be compared and contrasted. Entertainment employs simulation models for real-time interaction for user enjoyment.

The system analysis portion of M&S shares a unique relationship with Operations Research (OR). M&S and OR both are interdisciplinary areas of study that began in the 1930s. System analysis simulation traces its roots to the early 1930s with the documentation of the Monte Carlo method [2]. In the late 1930s OR scientist used the Monte Carlo method to evaluate certain wartime missions. OR employs mathematical models to produce optimal solutions to operations issues. M&S is often used to produce solutions that could not be solved analytically. There are many parallels and intertwining between OR and M&S. There are even those that contend that M&S is a subset of OR [3].

Pioneers of M&S like K.D. Tocher, Harry Markowitz and Robert Sargent were well known in the OR field. “Toch[er] was awarded the Silver Medal of the OR Society

in 1967, Honorary Fellowship of the British Computer Society in 1971, and was elected President of the OR Society for 1972–1973" [2]. He is also credited with the framework for discrete event simulation. Markowitz, an economist by training, was awarded the John von Neumann Theory Prize from the Operations Research Society of America for his work on portfolio theory, sparse matrix methods and SIMSCRIPT, a simulation programming language. Sargent was an OR professor at Syracuse University before joining the Electrical Engineering and Computer Science department. Sargent is renowned for his work on verification and validation of simulation models.

In order to understand the path on which M&S, particularly the system analysis portion of M&S, is traveling, this paper will explore its relationship with OR. This paper is organized in the manner that follows. The next section provides a brief history of Operations Research. This is followed by a section on the history of Modeling and Simulation. The fourth section provides an overview of the paradigms of OR and M&S. The paper concludes with a summary and thoughts of the future of Modeling and Simulation.

2. HISTORY OF OPERATIONS RESEARCH

Operations Research or Operational Research as it is known in Europe was born of necessity in World War II (WWII). British military gathered researchers from different disciplines to create a team designed to apply the scientific method to the allocation of scarce resources and the efficient use of the new radar tool [4]. The goal of this team was to provide research on operations to help decision makers make better decisions. The success of OR during the war efforts made it attractive to industry in the boom that followed WWII. Organizations with increasing complexity and specialization sought assistance from former military OR specialists [4]. OR gained widespread acceptance in academia, science and business by the mid-1960s [5].

Industry looked to scientific techniques to maximize the use of resources and profitability. In 1955 Dantzig introduced the simplex algorithm that is used to optimally solve a system of linear inequalities [6]. "The goal of the simplex method is to find how to use the available resources in the most profitable feasible way [4]." Dantzig notes that after presenting his linear programming problem, von Neumann conjectured the duality theorem. Duality theorem states that "every linear programming problem has associated with it another linear programming problem called the dual" [4]. While the primal is given in maximum form, the dual is given in minimal form. Von Neumann noted that it was an equivalent for his zero-sum game theory. This was later proven by Albert Tucker. The duality theorem has been used to design algorithms for combinatorial optimization problems [7].

As Operations Research expanded into areas like market research and management problems, not everyone accepted this work as a new discipline. Critics of OR conjectured that it was simply another application of scientific methods already well established. In 1954 John

Magee faced criticism when he attempted to describe Operation Research as an applied science [8]. His critic Richards wrote "that as procedures of investigation, neither operations research nor the scientific method ever can stand apart in the same epistemological sense as areas of academic or business research to which they are applied [9]." While this may be true, Operations Research has withstood the test of time.

The number of OR professionals grew and profession societies began to develop. The OR Club began in 1948 in London (UK) as the first OR association. The club held scientific meetings and produced the first quarterly OR journal. In 1957 the first International Federation of Operational Research Society (IFORS) was established in Oxford (UK). IFORS now has over 30,000 members in over 45 countries.[1]

Operations Research has developed into a science of optimization subject to constraints. OR boasts over 200 methods and techniques that practitioners can attempt to use for almost any problem [10]. This is both a strength and a weakness of OR. It provides efficiency in general activities providing a standard usable tool. However, it provides a false belief that any problem can be made to fit one of the techniques [10]. Furthermore, it is unlikely that anyone is an expert in, or even aware of, every tool available. This leads the analyst to specialize in one or two techniques and believe that these one or two techniques can solve any problem [3].

OR has focused on normative models to control the behavior of a system [10]. There is little accounting for or measurement of intangibles. However, most systems are socio-technical and rife with intangibles. "A socio-technical system is a collection of human and non-human parts interacting in an integrated way, in which overall system behavior arises from multiple cycles of interaction within and between the human and non-human parts" [11]. Saaty attempts to provide a structure to mathematically account for variables that are not typically value-based using an analytic hierarchy process [12]. Still OR as a whole does not acknowledge nor take into account non-instrumental, intrinsic value [5].

3. HISTORY OF MODELING AND SIMULATION

Unlike Operations Research, modeling and simulation cannot point to one specific event as the beginning of modeling and simulation. Simulation predates computers; artificial sampling to estimate π was documented in 1777 by Buffon [13]. Simulation had been used by scientists, mathematicians and OR professionals to solve problems that did not have an analytic solution. However, these simulations were often done by hand using tables of random numbers [2]. Computer simulations date back to the 1930s but the late 1950s into the 1960s brought significant advancement to M&S. During this time there were many advances in computing ability, software and thought.

[1] According to the IFORS website http://ifors.org/web/. Accessed on 12 March 2014.

IBM introduced the Fortran computer language in the late 1950s. Fortran moved programming away from the cryptic machine code required previously. Continued advances in computer software enhanced the use of simulation. The 1960s brought Algo160, a scientific language, and COBOL, a commercial and business oriented language into use. The 1960s also ushered in the second and third generation of computers. These systems were delivered with application software including simulation packages.

In 1957 MIT Professor Jay W. Forrester introduced the mathematical model known as System dynamics [14]. While OR was intended to optimize industrial activity, system dynamics (SD) was intended to show how organization structure, policies and timing of industrial activity interacted to influence the success of the enterprise [14]. SD modeled a system using differential equations to study the feedback loops. The simulation showed the aggregate effects over time.

That same year, K.D. Tocher, using a simulation package, was asked to design a simulation for the steel plants at United Steels. United Steels had multiple sites with different technologies. Tocher set out to create a comprehensive model that could be used at any of the sites. He envisioned a system of components that progressed through states that changed only at discrete events [2]. Tocher's General Steelplant Program was really a framework for General Simulation Program and is still the model of discrete-event simulation.

Schelling's Segregation Model in 1971 ushered in a new modeling platform called Agent-based Modeling and Simulation (ABMS). The ABMS framework provided a new method of modeling complex adaptive systems. Using an ABMS structure, Epstein and Axtell created an artificial society, Sugarscape, that demonstrated the ability of this platform examine social structure [15]. ABMS gave modelers a means to identify and investigate emergent behaviors.

4. PARADIGMS OF OR AND M&S

Mathematics has an axiomatic paradigm. It begins with statements deemed to be true –axioms. "No appeal to the real world is made, nor indeed held to be relevant. Theorems, or statements, then are deduced from these, which take their validity from the initial axioms" [16]. Physics and natural science are more empirical in their approach. Theories from the empirical paradigm are derived by testing a hypothesis that was developed from observed facts. There is, however, a third paradigm: constructivism. "The goal [of constructivism] is not to discover an existing truth, external to the actors involved in the process, but to construct a 'set of keys' which will open doors for the actors and allow them to proceed, to progress in accordance with their objectives and systems of value" [17]. It is within this realm that OR and M&S reside.

OR and M&S utilize models of systems that often cannot be empirically validated. The system either does not exist or cannot be manipulated to the state that is under study. The study is therefore

developed with a set of assumptions or elementary statements, deductive consequences and their relationships [16]. "Unfortunately, there is no commonly accepted set of assumptions usable for all complex system simulations" [18]. These statements are, therefore, judged on their 'reasonableness' with respect to the defined problem and the agreement of the parties involved. Valid assumptions for one problem may not necessarily be valid for another study [16]. This is one of the challenges of OR and M&S. Thus practitioners of OR and M&S typically take the instrumentalist view: the concern is not finding the truth but rather gaining insight into the problem at hand [3].

Defining the problem is critical to the ability to gain insight. Balci, from an M&S perspective, contends that to properly formulate the problem the modeler must first analyze the universe of discourse [19]. Whereas Landry et al. approach the situation from a managerial perspective and note that "a problem situation is very much defined and determined by the perception and behavior of the actors" [20]. Whether from the academic view of the discourse or the managerial view of behavior, the problem formation leads the modeler to the conceptual model. "The conceptual model is the coherent 'mental image' of the problem situation and is formed by the perceptions and value judgments of both model builders and decision-makers" [20]. The connection between model builder and decision-maker or model user is critical. If the problem that the builder solves is not the problem the user envisioned, the model will contain a Type III error – solving the wrong problem, and will be invalid to the user. Insight gained from a simulation will only be valuable if the parties involved agree that the model is valid [21].

Validation of models is a key component in OR and M&S. Sargent explains conceptual model validity as the underlying theories and assumptions are accurate, the problem representation is correct, and the structure, logic, mathematical and causal relationships are "reasonable" for the intended purpose [22]. This concept of validation is straightforward and easy to understand. However, it does not lead to a definitive procedure for model validation. There is no "general theory of model validation" in part because there is no agreement on what constitutes a valid model [23]. Nevertheless it is essential for OR and M&S to strive to create valid, usable models.

Usability is another important part of OR and M&S. "OR and M&S only exist to be applied to application problems…" [3]. Operations Research is often defined as the application of advanced analytical methods to help make better decisions. If it is not used in the decision making process, it serves no purpose. Likewise simulation models are intended to provide information about a system without requiring testing on the actual system. If a simulation model does not provide the user a beneficial comparison to the real system, it is useless. OR and simulation models are valuable and valid only if the user finds them credible. The usefulness of the model is the primary feature of the constructivist paradigm.

5. CONCLUSION

OR and M&S have paralleled and crossed paths many times in their more than 70 year existences. From Monte Carlo simulations to aid in WWII to OR analyst creating discrete-event simulations, M&S has contributed to the advancement of OR and OR has contributed to M&S. Both subjects call on multiple disciplines and both derive their value and validity from application more than theory. However, as the field of OR grew, OR scholars developed a disdain for M&S. "Many professionals in management science and operations research cast simulation as the 'method of last resort' and expressed the view that 'anyone could do it'" [1].

The discipline of operations research took the focus from formulating and solving problems in turbulent environments to almost an obsession with mathematical models and algorithms [5]. Yet even within the OR profession is the recognition "that the optimization tools so useful in industry were not adequate when applied to the problems of social systems…" [24]. The quantitative methods of the 1960s that allowed OR to flourish as a discipline could not address the more complex human systems.

M&S does not necessarily seek to solve problems with the same analytical precision as operations research (most likely the cause of the dissention). However, as Saaty notes, "It is better to solve the right problems approximately than to solve the wrong problem precisely" [10]. The challenge of M&S is to define the right problem. M&S draws heavily on the 'art of modeling' to define the problem. Problems in a socio-technical system are complex and dynamic therefore the models are also complex and dynamic. Modeling requires creativity and a deep understanding of the system being modeled. Verifying that the simulation performs as expected and proper analysis of the inputs and outputs are important but it is minor compared to designing a conceptual model that is a tool to enhance understanding of the system beyond the initial knowledge used to formulate the model [10].

While OR struggles with today's complexity, modeling and simulation struggles identifying its epistemology and ontology. The problems of M&S as a discipline are similar to those that OR faced in its early years. OR found its acceptance as a discipline by defining itself through its mathematical models and algorithms. However, in doing so, it lost some of its flexibility and pioneering spirit. As M&S continues to define its precepts, it is incumbent on the community to ensure that the flexibility and usefulness is not lost.

REFERENCES

[1] R. E. Nance and R. G. Sargent, "Perspectives on the evolution of simulation," *Operations Research,* vol. 50, pp. 161-172, 2002.

[2] B. W. Hollocks, "Intelligence, innovation and integrity—KD Tocher and the dawn of simulation," *Journal of Simulation,* vol. 2, pp. 128-137, 2008.

[3] A. Collins, "Which is Worse: Large-Scale Simulations or the 80% Solution?," *SCS M&S Magazine,* pp. 27-38, 2012.

[4] F. S. Hillier and G. J. Lieberman, *Introduction to Operations Research*, Eighth ed. New York, NY: McGraw Hill, 2005.

[5] R. L. Ackoff, "The Future of Operational-Research Is Past," *General Systems,* vol. 24, pp. 241-252, 1979.
[6] G. B. Dantzig, A. Orden, and P. Wolfe, "The generalized simplex method for minimizing a linear form under linear inequality restraints," *Pacific Journal of Mathematics,* vol. 5, pp. 183-195, 1955.
[7] M. X. Goemans and D. P. Williamson, "The primal-dual method for approximation algorithms and its application to network design problems," *Approximation algorithms for NP-hard problems,* pp. 144-191, 1997.
[8] J. F. Magee, "Application of Operations Research to Marketing and Related Management Problems," *Journal of Marketing,* vol. 18, pp. 361-369, Apr 1954.
[9] E. A. Richards and J. F. Magee, "Operations Research or The Scientific Method," *The Journal of Marketing,* pp. 159-161, 1954.
[10] T. L. Saaty, "Reflections and projections on creativity in operations research and management science: a pressing need for a shift in paradigm," *Operations Research,* vol. 46, pp. 9-16, 1998.
[11] G. L. Mathieson and K. A. Richardson, *Knots, Lace and Tartan: Making Sense of Complex Human Systems in Military Operations Research: the Selected Works of Graham L. Mathieson*: ISCE Pub., 2009.
[12] T. L. Saaty, "How to make a decision: the analytic hierarchy process," *European journal of operational research,* vol. 48, pp. 9-26, 1990.
[13] B. Jansson, "Random number generators," 1966.
[14] J. W. Forrester, "Industrial dynamics: a major breakthrough for decision makers," *Harvard business review,* vol. 36, pp. 37-66, 1958.
[15] J. M. Epstein and R. Axtell, "Artificial societies and generative social science," *Artificial Life and Robotics,* vol. 1, pp. 33-34, 1997.
[16] W. J. Strauss, "The Nature and Validity of Operations-Research Studies, with Emphasis on Force Composition," *Operations Research,* vol. 8, pp. 675-693, 1960.
[17] B. Roy, "Decision science or decision-aid science?," *European journal of operational research,* vol. 66, pp. 184-203, 1993.
[18] A. Tolk, *Ontology, Epistemology, and Teleology for Modeling and Simulation: Philosophical Foundations for Intelligent M&S Applications* vol. 44: Springer, 2012.
[19] O. Balci and W. Ormsby, "Conceptual modelling for designing large-scale simulations," *Journal of Simulation,* vol. 1, pp. 175-186, 2007.
[20] M. Landry, J.-L. Malouin, and M. Oral, "Model validation in operations research," *European Journal of Operational Research,* vol. 14, pp. 207-220, 1983.
[21] M. Pidd, *Tools for thinking*: Wiley Chichester, 2003.
[22] R. G. Sargent, "Verification and validation of simulation models," in *Proceedings of the 37th conference on Winter simulation*, 2005, pp. 130-143.
[23] M. Landry and M. Oral, "In search of a valid view of model validation for operations research," *European Journal of Operational Research,* vol. 66, pp. 161-167, 1993.
[24] K. Tocher, "The dilemmas of operational research," *Operational Research Quarterly,* pp. 105-115, 1972.

High-Order and Attributed Motifs in Complex Networks

David Wright, Old Dominion University, MSVE Department
dwrig032@odu.edu

Abstract- Social networks often display a broad range of structures and functions. One interesting set of structures at the mesoscopic scale, network motifs, has been studied thoroughly in the last decade. The earliest work [Milo et al. 2002] identified the structure of the simplest motifs and hypothesized their function, and increasingly more efficient algorithms for network motif discovery have been presented [Kashtan et al., 2004, Schreiber et al., 2005, Wernicke, 2006]. However, the motifs studied above are static topological structures, often without overlaid functionality, and the merging of attributes and motifs has received little attention. To address this issue, the following work details the collection and analysis of massive amounts of data from a diverse set of networks, and describes and compares structural patterns of attributed motifs at multiple levels of network hierarchy, called *high-order motifs*.

Introduction

Social and information network analysis provides insight into the structure and dynamics of complex networks. When people interact on the web, they express themselves, typically through adding or removing links to other people. However, studies of network motifs tend to focus on unattributed graphs—that is, the edges of the graphs contained only one type, such as positive or negative, or neutral. Graphs that have attributed edges are far more expressive and convey much more information about the users and the resulting network. For instance, in a graph with edges that can be positive or negative, a positive edge can convey trust, approval, or fondness, while a negative edge reveals distrust, disapproval, or dislike.

The main focus of this work is to explore the significance of subgraphs with rich edge data, extending previous work on motifs that overwhelmingly focus on small graphs with homogeneous edges—that is, edges were unattributed, with a few notable exceptions [Lescovec et al., 2010a,b].

This research explores the structure and evolution of networks from the point of view of network motifs. How are networks configured? How do they evolve over time? What features are common among networks from diverse domains, and what features are different? What recurrent patterns are found in networks? Can we make predictions about networks? Does sentiment flow over networks, and are there motifs of sentiment? Uncovering the statistical properties of network motifs provides insight into problems such as link prediction, community dynamics, information cascades, influence maximization, and neural network model selection.

Background

There are several metrics used to analyze real-world networks: diameter, clustering coefficient, degree distribution, and the presence or absence of hubs. Social networks display low network diameter (and, after a startup period, even shrinking diameter [Leskovec et al, 2005]), high clustering coefficient, power-law degree distribution, and the presence of hubs or authorities. Biological networks—protein-protein interaction networks in particular—display similar properties.

Real-world networks come in many flavors: co-authorship (journal and conference publications), transportation (roads, subways), infrastructure (water pipes, electrical grid), biological (protein interaction, RNA transcription), technological (WWW and internet), social (Twitter, Facebook), affiliation (authors-to-conferences, customer-to-product), information (product co-purchasing). This list is not exhaustive.

One of the first results from random graph theory [Erdos and Renyi, 1959] was calculating the conditions under which subgraphs are statistically guaranteed to appear in a random graph. These random graphs typically do not reflect real-world networks and are therefore used as null models in network hypothesis testing. Since [Milo et al., 2004] first described motifs and their existence in several diverse domains—cell biology, electrical circuits, food webs—there have been numerous investigations of motifs in complex networks.

Motifs. Network *motifs*—connected subgraph of a few nodes—are a kind of building block for complex networks, and were first identified in biological networks [Milo et al. 2002], specifically protein-protein interaction (PPI) networks. Similar motifs—also known as *graphlets* or *subgraphs*—were also identified by Milo et al. in a broad range of networks: transcriptional, food web, neuronal, electrical circuit, WWW, and internet. Further studies have found network motifs in transportation [Jiang et al., 2005], citation [Conway, 2011], collaboration [Krumov et al., 2011], neural [Sporns and Kötter, 2004], and ecological [Stouffer et al., 2007]. Not only do all types of networks above have motifs, but certain motifs are significant in more than one type of network. For example, a feed-forward loop is found in networks of gene regulation, neurons, and electrical circuits. The figure below shows all unique connected motifs of size 3 where the edges are directed and have no attributes. Motif 5 below is a feed-forward loop. Motif 7 is known as a feed-back loop. Motif 11 is called an up-linked mutual dyad.

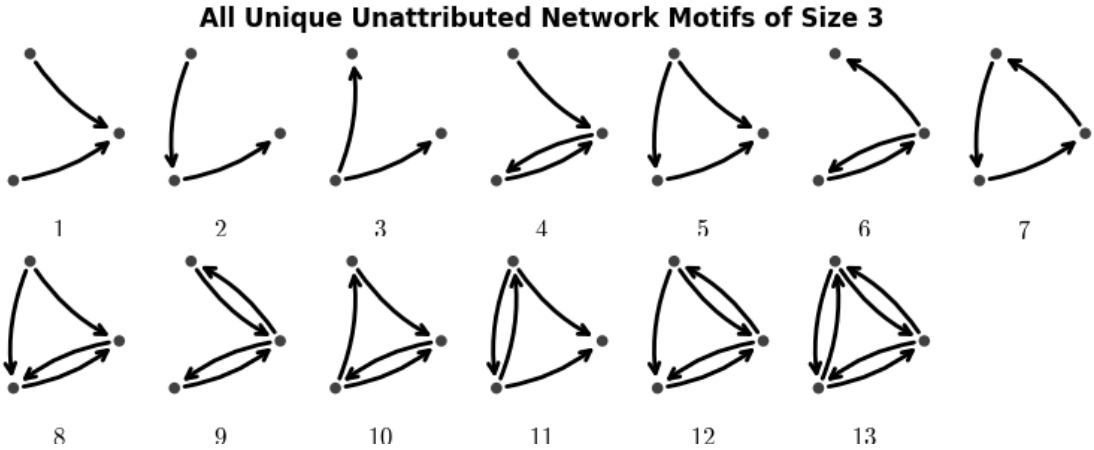


Figure 1

The figure above represents all unique motifs—that is, there are other motifs not shown because they are isomorphic to a motif already included. Below are all directed unattributed network motifs of size 3. The 13 motifs above are representative of the 54 motifs below. For example, notice that motifs 2, 3, 5, 7, 10, and 11 below are the same as motif 2 above. Hence motif 2 above has an isomorphic set of size 6.

Most previous work focuses on motifs at the microscopic network level. Several exceptions exist. [Song et al., 2005] use a *box covering* method to hierarchically collapse a network into one single node. Also, the work of [Leskovec et al., 2005] shows that networks can be represented recursively using very few parameters using the method of *Kronecker multiplication*, a tensor matrix operation. This dissertation studies motifs at all three network levels: microscopic, mesoscopic, and macroscopic. By looking at network motifs at different levels of network resolution, we gain better structural knowledge of the network.

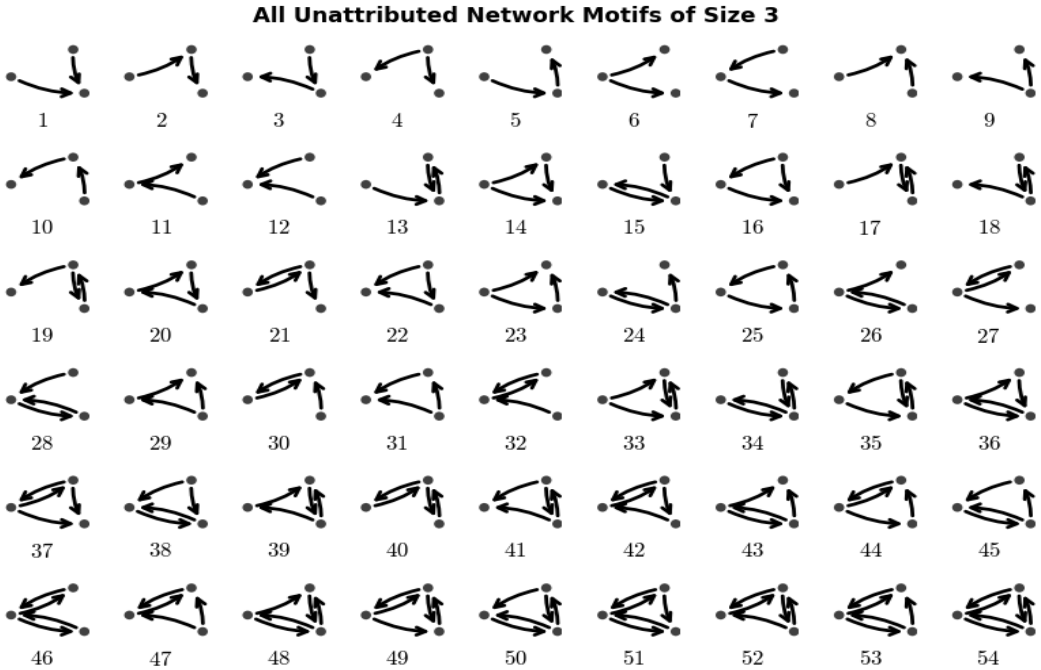


Figure 2

Attributed Motifs. The generative network algorithms delineated above generate graphs that resemble real-world graphs quite well in the mechanistic sense—that is, the synthetic graphs match various metrics such as small network diameter, high clustering coefficients, and power-law degree distributions. But the generative models often only account for structural properties of networks, often overlooking the attributive nature of real-world networks. For example, each

member of a social network has numerous attributes, including demographic and preferential data; these are called node *attributes*. Edges can have attributes as well. For example, if the edge between two nodes represents an email, the edge attributes include a timestamp, the length and language of the text, any photo attachments, the sentiment or topic, and so on.

Limited studies have been published that directly merge the concepts of network motifs and node attributes. [Lescovec et al., 2010a,b] study motifs—*triads* in their paper—of trust in Epinions, Slashdot, and Wikipedia datasets. They limit the study to dyads and motifs with three nodes and only one edge between each pair of nodes. Their motivating application is signed link prediction. Results include evidence for the theories of balance and status, the latter having stronger evidence. Balance theory is limited to undirected graphs and states that 'the friend of a friend is a friend,' and 'the enemy of a friend is an enemy' and so on. Status theory operates on directed graphs, and assumes that the direction and sign of edges are predictive attributes. The authors present a logistic regression model to predict a positive edge. The feature vector of this model is composed of degree of the nodes involved and participation in motifs. This method will be fully explained in the section 3: Methodology. The table below gives, for directed and undirected graphs, the number of unique motifs of different sizes, *k*, and for different numbers of possible edge attributes. The numbers grow exponentially.

Table 1.

directed					
attributes	**1**	**2**	**3**	**4**	**5**
k=2	2	5	9	14	20
k=3	13	132	710	2660	7875

undirected					
attributes	**1**	**2**	**3**	**4**	**5**
k=2	1	2	3	4	5
k=3	2	7	16	30	50
k=4	6	53	250	855	2380

Another close work is that of [Kim et al., 2013] on Multiplicative Attribute Graph (MAG) models, which models attributed nodes and gives link predictions as a function of attribute similarity. [Chechik et al., 2008] describes *activity motifs*, or motifs of network dynamics; specifically, two extensions to the standard network motif paradigm are given: *timing* activity motifs, and *binding* activity motifs. Their work deals with biological cellular processes, where they discover the existence of timed motifs in biochemical chains of events—where events are represented by a combination of nodes and edges overlaid with functional data—such as gene activation. For example, there are three cellular events *X*, *Y*, and *Z*, which occur in a particular order much more frequently than they would in random order, e.g. *XXYZ*. This pattern on any number of nodes $n > 1$ is known as a forward activation chain, where the order of events in the string *XXYZ* is also the temporal order in which they occur; there is also a backward activation chain in which the order is reversed, e.g. the order in which the above activity motif XXYZ occurs is actually *Z*, *Y*, *X*, *X*.

In addition to the two chains motifs described above, [Chechik et al. 2008] also classifies branching activity motifs, allowing for forking and funneling topology, with events firing sequentially and/or in parallel. In contrast to structural motifs, which deal with nodes and edges as a static structure, the timing activity motifs deal with an order of events. Also of interest is [Shafiq et al., 2013], which models the morphology of cascades, which are directed, acyclic motifs—trees.

The table below shows all unique attributed motifs of size three, where there are two choices for an edge attribute: positive or negative.

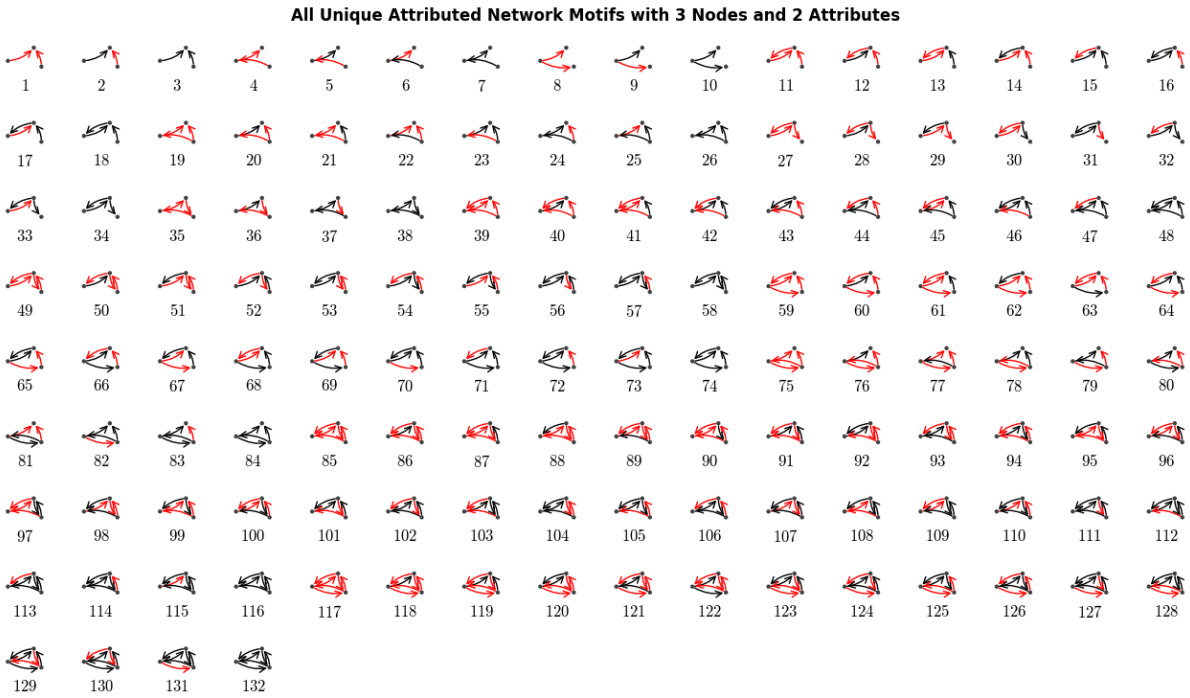


Figure 3. All 132 unique connected motifs of size k=3 nodes, where there are two choices of attribute.

Dataset Description

Two datasets are used in this study: Epinions and Slashdot. Both networks display power-law degree distributions, small diameter, and high clustering.

Epinions is a website that allows users to review products and other reviewers. The dataset contains 131,828 nodes and 841,372 edges. The graph amounts to a who-trusts-whom network. Positive edges account for roughly 85% of all edges; 15% are negative.

Slashdot is a user-curated news website. The *Slashdot Zoo* feature allows users to specify whether another user is a friend or a foe. These labels are mapped to the set {-1, +1}, where friend is represented by +1 and foe is -1. Labels are based on users' reviews and comments on news articles. If a user A likes the comments posted by another user B, they user A will add a link from A to B with edge sign +1; alternatively, if user A does not appreciate the comments left by user B, then user A can add a disapproving -1 edge from A to B. The Slashdot dataset contains 82,140 nodes and 549,202 edges. Positive edges account for 77.4% of edges and negative edges account for 22.6%.

Below is a plot of the node degree distribution of the Epinions and Slashdot network. As we can see, both follow a power-law degree distribution. In fact, the α components of the two networks are quite close: 1.60 for the Epinions network and 1.49 for the Slashdot network.

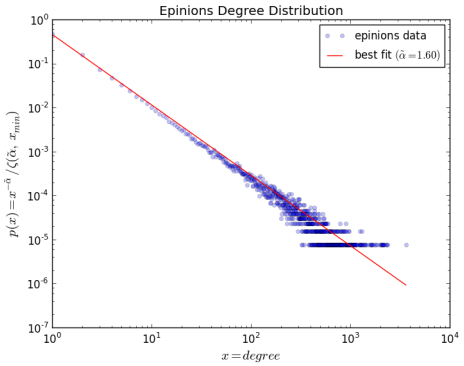


Figure 4

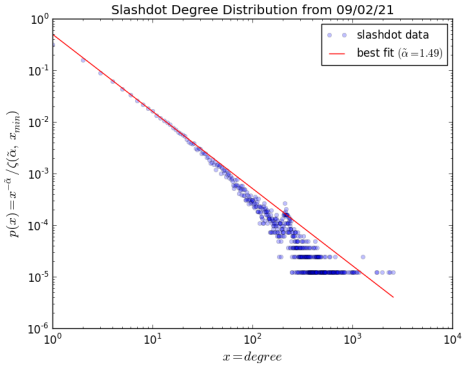


Figure 5

Methodology

This section will discuss the various motif discovery algorithms, z-scores for quantifying the significance of particular motifs, and current graph generation methods. Motif discovery involves three steps: first, the subgraph must be counted in the graph of interest; second, the corresponding subgraph must be counted from a null model; finally, the two previous counts must be compared. Various choices must be made for this process to be valid, such as choice of null model, and the counting method. There are also memory and cpu issues that limit the depth of the study.

Null models. A motif is considered significant in a real-world network if it occurs more frequently than in a null model. In related research, a null model of choice is the random graph produced by the Erdos-Renyi model, as described in section 2.4. This model is typically generated to match various metrics while leaving others to chance. For example, the number of nodes n and edges m of the null graph $G_{n,m}$ can be matched to a real-world network, but the degree distribution will be binomial—as opposed to heavy-tailed, a distribution found frequently in the wild—because of the particular method of graph construction. Equivalent to matching nodes and edges is matching nodes n and the probability p of an edge, denoted $G_{n,p}$.

The two previous examples give binomial degree distributions, but the null model will be more valid if the degree sequence also matches—real-world social networks are rarely governed by a binomial degree distribution. The work of [Milo et al. 2004b] gives algorithms for generating synthetic graphs with prescribed degree distributions. The configuration model [Newman et al., 2001, Newman, 2003] generates random graphs with a given degree sequence, and includes the possibility of self-loops, which are not always valid.

An alternative to the null graphs presented above is one that uses the exact edges and nodes of the real-world model, but randomly reassigns the edge or node attributes. For example, if we use this method to generate a null model for the signed Epinions network, where each edge is directed and has an attribute $X \in \{-1, +1\}$, the degree distribution of the null model is exactly the same as the real-world network, but the $-1's$ and $+1's$ are randomly permuted. Therefore the synthetic model matches the real-world network in degree distribution, but is null in attribute distribution. This type of shuffling is used in [Lescovec et al., 2010a].

There are several ways to achieve the attribute shuffling. The first method involves empirically determining the probabilities of the edge values by counting the occurrences of each value and dividing by the total number of values. Assuming, for the null model, the edge values are independent, the likelihood of occurrence of -1 and $+1$ follows a distribution $X \sim \text{Bernoulli}(p)$. For the Epinions network, $p = Pr(X = +1) = 0.85$ and $q = Pr(X = -1) = 1 - p = 0.15$. Once we know the proportions, we know p and q, and we can iterate through the edges in the graph and for each edge assign a weight randomly from the attribute distribution X, resulting in a graph with the expected number of each attribute equal to that of the real-world graph. Since we are using the exact node-to-edge structural relationship as the original model, we do not have to generate the underlying graph structure—the edge set E—and since we are only performing $|E|$ Bernoulli trials, the algorithm runs in $O(|E|)$ time and memory (if an adjacency list data structure is used to represent the network). The resulting distribution of the attributes in a is $Y = \sum_{k=1}^{n} X_k \sim \text{Binomial}(n, p)$.

The second method involves generating a list of the attributes from the original graph that is sampled (with replacement) uniformly at random for each edge of the null graph, resulting in the same expected value as the algorithm above.

The above algorithms do not guarantee that the original counts of $-1's$ and $+1's$ will be the same; in expectation, they are the same, but they often differ in the resulting null model. There are two ways to preserve the edge value counts. In order to keep the exact number of $-1's$ and $+1's$, we can generate a list of the edge values as above, then permute or shuffle the list, and redistribute the values to the same edges. Alternatively, we can simply sample without replacement from a list of the original edge values.

These null graphs are used for comparison to real-world graphs. This dissertation will use the counts of motifs in null graphs as a baseline for comparison with counts of motifs in real-world networks. The process of comparing and scoring motifs is described below.

Motif discovery algorithms. The discovery of motifs in biological networks was due to the work in [Milo et al., 2002], which uses a brute-force algorithm that performs well for small motifs, up to size 4. Due to the inherent difficultly in enumerating subgraph of larger sizes, various heuristics have been proposed. Hence there are two flavors of motif discovery algorithm: exact, and estimated. For the large graphs studied here, a heuristic approach was used. The approach is as follows. Nodes are sampled uniformly at random, and the motifs of which they are a part are counted. These counts are aggregated among all nodes sampled. 100 nodes are sampled.

Motif significance. A motif is considered *significant* in a real-world network if it occurs more frequently than it would in a random graph. A motif G' of a graph G is considered *significant* or *recurrent* if its frequency $F_G(G')$ is above some predefined threshold. There are two main methods of determining whether a subgraph is statistically significant: the Z-score and the P-value.

The Z-score defined in [Milo et al., 2002] uses the frequency of the motif in the real-world graph, $F_G(G')$, the mean frequency $\mu_R(G')$ of the motif G' in N random graphs R_i, where $1 \leq i \leq N$, and the standard deviation frequency, $\sigma_R(G')$ of the same set R. The final equation is

$$Z(G') = \frac{F_G(G') - \mu_R(G')}{\sigma_R(G')}$$

R is the set of all N random graphs. Motifs with high Z-scores are considered statistically significant. This research will study various facets of motifs, including motifs at all levels of hierarchy—that is, motifs of nodes, of communities, of communities of communities, and so on—the flow of emotion or sentiment through motifs, the temporal dynamics of

motifs, and the inclination of nodes to occupy certain positions on a motif.

The above equation works for small graphs where the actual counts can be obtained efficiently. However, for very large graphs, sampling must be performed. Therefore the $F_G(G')$ term is replaced by $\mu_G(G')$, the average count of motif G' from N samples of the graph. Specifically, the graph is sampled N times, and for each sample, M nodes are sampled. For each node sampled, we count at most 100 motifs associated with that node. This results in the following equation:

$$Z(G') = \frac{\mu_G(G') - \mu_R(G')}{\sigma_R(G')}$$

An alternative method of identifying significant motifs comes from [Lescovec et al., 2010a] in the form of a *surprise* factor, which is defined as the number of standard deviations between the actual occurrences of a motif and the expected number of occurrences after the edge signs have been shuffled.

High-order motifs. High-order motifs are motifs at higher levels of network structure, such as the community level. Motifs from the literature are motifs on individual nodes and their edges. But there has been little investigation into inter-community motifs. The individual node level is level 0. Using the partitioning methods given by [Newman, 2006], the network can be partitioned into communities; these communities combined with the edges that go between them give motifs of level 1.

Several decisions must be made regarding the construction of high-order motifs. In this dissertation, the values of the edges between communities—the values of the node-node edges at level 0 whose two endpoints are in two different communities—are aggregated and the final value given to the inter-community edge is the most likely value from the aggregated list. For example, if the edges between two communities have values [+1, +1, -1, +1], then the edge is given the value +1.

Using motifs at all levels 0, 1, …, L-1 where L is the level resulting in one high-order community—that is, one node—we can compare motif significance at different levels, and determine the fractal nature, if any, of the network with regard to motifs. We can also examine the differences between communities. Are there patterns to inter-community motifs or signed edges? Do different networks display similar motif significance vectors? If so, at how many levels does this similarity hold?

Results

The somewhat small size of the network graphs limits the study of high-order motifs to two levels: node level, and community level. There are not enough nodes at level three for meaningful analysis. Thus the plots below have two lines each. There are, for some motifs, close z-scores for both levels. For example, there are several motifs for which the z-score spikes up (or down) at both levels of hierarchy. This result implies that there is a fractal nature to motif structure; meaning small graph structures are apparent and significant at the inter-node level and at the inter-community level. Not only are the motifs similar for the same graph, they are similar in both networks. Observing the two plots together shows that most of the significant motifs in the Epinions network are also significant in the Slashdot network.

These results are artifacts of both the network and the analysis, and certain decisions could have biased the results. The selection of graph partitioning algorithm could affect the outcome by favoring certain types of clusters or communities whose makeup is inclined to behave in a particular way. Also, the use of the *mode* in selecting inter-community edge signs could have biased the results.

Conclusions

In this paper, we demonstrated the existence of significant motifs in signed social networks. Further, the existence of motifs generalizes across different levels of network hierarchy—that is, intra-network generalization—and across different datasets—inter-network generalization.

Future work will involve validation on more directed datasets, such as the Wikipedia Administrator Voting network, and will also involve undirected networks. The undirected networks—if large enough—will be useful in determining similarity at levels above two. We will also apply signed motifs to the problem of link prediction, where we know a link forms and we must predict the sign of the link. We will use machine learning to build a model, the features of which will include detailed data about a node's motif participation, in both the *id* of the motif and the location of the node in the motif. This extends previous work [Lescovec et al., 2010b]. The hypothesis is that knowning the location of a node in a motif tells much more than just knowing if a node is *somewhere* in the motif. This will undoubtedly increase the size of the feature vector, but not beyond computational feasibility.

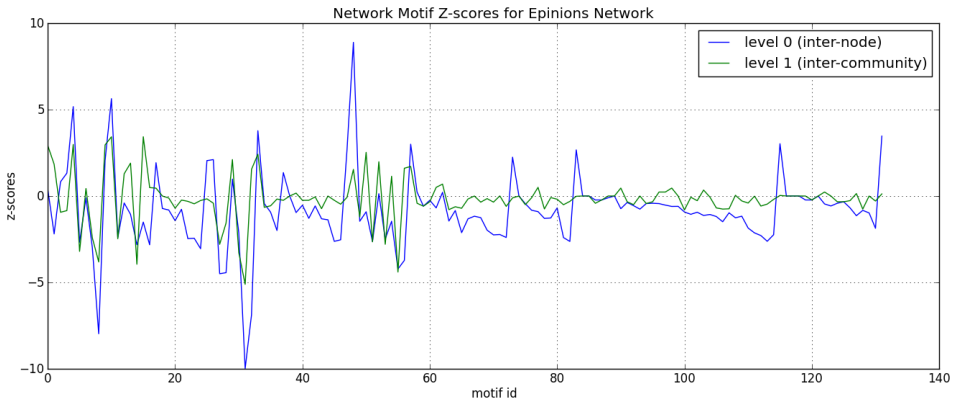


Figure 6. z-scores for motifs in the Epinions network.

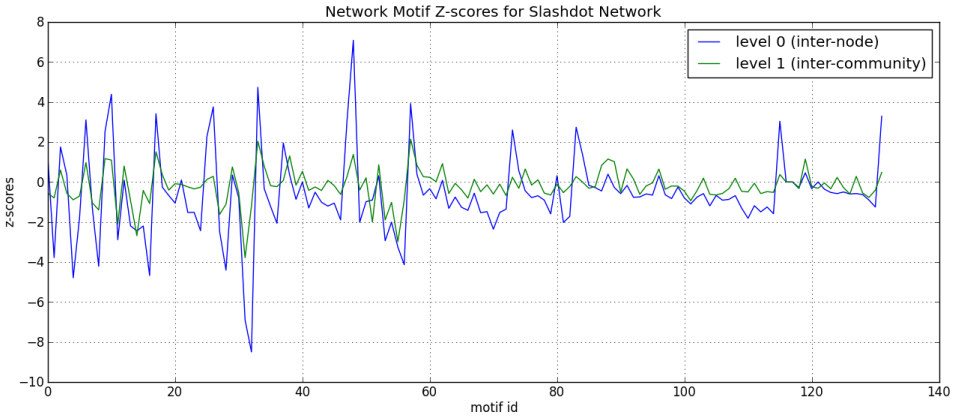


Figure 7. z-scores for motifs in the Slashdot network.

References

R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics,* 74 (1):47–97, 2002

A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science,* 286:509–512, 1999.

Bryden, John and Funk, Sebastian and Jansen, Vincent. Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science* 2013, **2**:3. doi:10.1140/epjds15.

Chechik G, Oh E, Rando O, Weissman J, Regev A, Koller D (November 2008). "Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network". *Nat. Biotechnol.* **26** (11): 1251–9

Conway, D. (2011). Modeling Network Evolution Using Graph Motifs. eprint arXiv:1105.0902

C. Cooper and A. Frieze. A general model of web graphs. *Random Structures and Algorithms,* 22(3): 311–335, 2003.

D.J. de Solla Price. "A general theory of bibliometric and other cumulative advantage processes". *Journal of the American Society for Information Science* **27**: 292–306. 1976. doi:10.1002/asi.4630270505.

P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae* **6**: 290–297. 1959.

P. Erdos and A. Renyi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Acadamy of Science,* 5:17–67, 1960.

Jiang, W., Vaidya, J., Balaporia, Z., Clifton, C., Banich, B. (2005) Knowledge discovery from transportation network data. ICDE 2005. Proceedings. 21st International Conference on Data Engineering.

Kashtan N, Itzkovitz S, Milo R, Alon U (2004). "Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs". *Bioinformatics* **20** (11): 1746–1758.

J. M. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models and methods. In *COCOON '99: Proceedings of the International Conference on Combinatorics and Computing,* 1999.

L. Krumov, C. Fretter, M. Muller-Hannemann, K. Weihe, and M.T. Hutt. Motifs in co-authorship networks and their relation to the impact of scientific publications. Eur. Phys. J. B 84, 535–540 (2011).

J. Leskovec, D. Chakrabarti, J. M. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *PKDD '05: Proceedings of the 9th Eu- ropean*

Conference on Principles and Practice of Knowledge Discovery in Databases, pages 133–145, 2005.

J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning,* 2007.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme- tracking and the dynamics of the news cycle. In *Proceed- ings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09,* 497.

J. Leskovec, D. Huttenlocher, J. Kleinberg. Signed Networks in Social Media. CHI 2010.

J. Leskovec, D. Huttenlocher, J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. WWW 2010.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (October 2002). "Network motifs: simple building blocks of complex networks". *Science* **298** (5594): 824–7

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, Uri Alon. (March 2004). "Superfamilies of Evolved and Designed Networks." Science 303. 1538-1542.

R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *ArXiv,* cond-mat/0312028, May 2004.

Newman, M. E. J. and Strogatz, S. H. and Watts, D. J. Random graphs with arbitrary degree distributions and their applications Phys. Rev. E, 64, 026118 (2001)

M.E.J. Newman, "The structure and function of complex networks", SIAM REVIEW 45-2, pp 167-256, 2003

M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, December 2005.

Newman, M. E. J. (2006). "Modularity and community structure in networks". *PNAS* **103** (23): 8577–8696.doi:10.1073/pnas.0601602103. PMC 1482622. PMID 16723398.

Schreiber F, Schwöbbermeyer H (2005). "Frequency concepts and pattern detection for the analysis of motifs in networks". *Transactions on Computational Systems Biology III*: 89–104.

Shafiq, M. Zubair, and Alex X. Liu. "Modeling Morphology of Social Network Cascades." *arXiv preprint arXiv:1302.2376* (2013).

Chaoming Song, Shlomo Havlin, Hernan A. Makse. Self-similarity of complex networks. Nature, 433, (2005), 392-395.

H. A. Simon. "On a Class of Skew Distribution Functions". *Biometrika* **42** (3–4): 425–440. December, 1955. doi:10.1093/biomet/42.3-4.425.

Sporns, O, Kötter, R (2004) Motifs in Brain Networks. PLoS Biol 2(11): e369. doi: 10.1371/journal.pbio.0020369

Sporns, O. (2006) Small-world connectivity, motif composition, and complexity of fractal neuronal connections. BioSystems 85, 55–64.

Stouffer, D., Camacho, J., Wenxin, J., and Amaral, L.A.N. Evidence for the existence of a robust pattern of prey selection in food webs. Proc. R. Soc. B. 22 August, 2007, vol. 274, no. 1621, 1931-1940.

Duncan Watts & Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature* 393, 440-442. 4 June 1998.

Wernicke S (2006). "Efficient detection of network motifs". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3** (4): 347–359.

Udny Yule. "A Mathematical Theory of Evolution Based on the Conclusions of Dr. J. C. Willis, F.R.S.". *Journal of the Royal Statistical Society* **88** (3): 433–436. 1925. doi:10.2307/2341419. JSTOR 2341419.

G. K. Zipf. *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, Massachusetts, 1949.

GPU Accelerated Randomized Singular Value Decomposition and Its Application in Image Compression

Hao Ji and Yaohang Li
Department of Computer Science
Old Dominion University
hji@cs.odu.edu, yaohang@cs.odu.edu

Abstract
In this paper, we present a GPU-accelerated implementation of randomized Singular Value Decomposition (SVD) algorithm on a large matrix to rapidly approximate the top-k dominating singular values and correspondent singular vectors. The fundamental idea of randomized SVD is to condense a large matrix into a small dense matrix by random sampling while keeping the important information. Then performing traditional deterministic SVD on this small dense matrix reveals the top- k dominating singular values/singular vectors approximation. The randomized SVD algorithm is suitable for the GPU architecture; however, our study finds that the key bottleneck lies on the SVD computation of the small matrix. Our solution is to modify the randomized SVD algorithm by applying SVD to a derived small square matrix instead as well as a hybrid GPU-CPU scheme. Our GPU-accelerated randomized SVD implementation is around 6~7 times faster than the corresponding CPU version. Our experimental results demonstrate that the GPU-accelerated randomized SVD implementation can be effectively used in image compression.

Keywords: Random Sampling, Singular Value Decomposition, Low-Rank Approximation, Image Compression, Graphics Processing Unit

1. Introduction

A factorization of a real matrix $A \in \mathbb{R}^{m*n}$ is singular value decomposition (SVD) if

$$A = U * \Sigma * V^T$$

where $U \in \mathbb{R}^{m*m}$ and $V \in \mathbb{R}^{n*n}$ are matrices with orthonormal columns, $\Sigma \in \mathbb{R}^{m*n}$ is a diagonal matrix whose elements, $\sigma_1, \sigma_2, \dots, \sigma_n$, are nonnegative singular values in non-decreasing order. SVD plays an important role in a wide range of modeling and simulation applications, such as modeling of genome-wide expression data [1], large-scale atmosphere-ocean interaction analysis [2], data-unfolding in high energy physics [3], recommendation engine in social network modeling [4] and simulations with MRI data [5], etc. One primary advantage of using SVD is that the low rank approximation A_k to matrix A with rank k can be readily formed as

$$A_k = M * N$$

where M is an $m * k$ matrix and N is an $k * n$ matrix. Consequently, the factorized matrices containing most important characteristics of the original matrix can be used for efficient modeling and computing, while those small matrices are inexpensive to store and manipulate.

The traditional deterministic SVD algorithm [6] on a large matrix is computationally intensive, which has cubic-time complexity with respect to the size of the given matrix. For an $m * n$ matrix A, when both m and n are large, deterministic SVD also requires large memory space. Randomized SVD [7-10, 15], by contrast, offers efficient alternatives to approximate the dominant singular components. Williams and Seeger [7] proposed a Nyström method to accelerate decomposition in kernel machines. Frieze et al. [8] studied column-sampling method for finding low-rank approximations in constant time. Drineas [9] modified column-sampling method to make it fit the pass-efficient model of data-streaming computation. Holmes et al. [15] developed a cosine tree sampling method for fast approximation of the complete matrix SVD. Halko et al. [10] performed random sampling on the original matrix to construct a small condensed subspace. The dominant actions of the original matrix A could be quickly estimated from this small subspace with relatively low computation cost and high confidence. Matrix operations on the projected small subspace allow randomized SVD algorithms to take advantage of the emerging high performance computing platforms, for instance, distributed memory systems, multi-core processors, multi-general purpose graphics process units (GPGPU), and the Cloud computing infrastructure [11, 14].

Graphics Processing Unit (GPU) is a specialized single-chip processor to take advantage of parallelism to achieve high performance computing. Many high performance linear algebra libraries on GPU architectures are available. For instance, CUBLAS (CUDA Basic Linear Algebra Subroutines) [12] contains the GPU-accelerated functions of basic dense matrix operations. Complementary to CUBLAS, CULA [13] is an extended linear algebra library provides high-level equivalent routines of LAPACK over CUDA runtime. Based on these GPU-accelerated libraries, Foster et al. [16] designed a GPU-based cosine tree sampling algorithm for

column-sampling SVD and achieved speedup of 6~7 over CPU implementation in large matrices.

In this paper, we focus on the randomized SVD algorithm proposed by Halko et al. [10] and present a GPU-accelerated implementation to quickly obtain the approximate of dominant singular components of a given large matrix. We find that the main bottleneck in the GPU implementation is the deterministic SVD on GPU with "short-and-wide" matrix. Using SVD decomposition on a derived square matrix instead can significantly reduce the overall computational time. In addition, in the case of matrices with a small dominant rank k value, if a hybrid GPU-CPU scheme is carried out, the efficiency of our implementation can be further improved.

The rest of the paper is organized as follows. Section 2 describes the randomized SVD algorithm. In section 3, we analyze the GPU-accelerated implementation of randomized SVD algorithm. Section 4 shows experimental results on large matrices and a NASA Mars image. Section 5 summarizes the paper.

2. The Randomized SVD Algorithm

The Randomized SVD algorithm was introduced by Halko et al. [10-11] to obtain a low-rank approximation of a large matrix. Instead of directly performing deterministic SVD on a large matrix, which is usually not only computationally costly but also memory intensive, the randomized SVD algorithm starts from a small random subspace and then projects the original matrix onto this subspace. The fundamental idea is, as the most important characteristics of the original matrix A are condensed into a small randomized subspace, this projected subspace becomes an amenable choice to approximate matrix decomposition but avoiding high computational cost. Algorithm 1 describes the Randomized SVD algorithm given an input matrix.

Algorithm 1 Randomized SVD

Input: $A \in \mathbb{R}^{m*n}$, $k \in \mathbb{N}$ and $p \in \mathbb{N}$ satisfying $k + p \leq min(m, n)$.

Output: k-rank randomized SVD components $U \in \mathbb{R}^{m*k}$, $\Sigma \in \mathbb{R}^{k*k}$ and $V \in \mathbb{R}^{k*n}$

1. Construct an $n * (k + p)$ random matrix Ω
2. $Y = A\Omega$
3. Compute an orthogonal basis $Q = qr(Y)$
4. $B = Q^T A$
5. $[U_B, \Sigma_B, V_B] = svd(B)$
6. Update $U_B = Q U_B$
7. $U = U_B(:, 1:k), \Sigma = \Sigma_B(1:k, 1:k)$ and $V = V_B(:, 1:k)$

To show how random sampling of A can extract information of the top k singular values/vectors, let $\omega_j \in \mathbb{R}^n$ and $y_j \in \mathbb{R}^n$ denote the jth column vector of random matrix Ω and the jth column vector of matrix Y, respectively. Since each element in Ω is chosen independently, ω_j can be represented as

$$\omega_j = c_{1j} v_1 + c_{2j} v_2 + \cdots + c_{nj} v_n, \quad j = 1, \ldots, k + p$$

where $v_i \in \mathbb{R}^n$ is ith right singular vector of matrix A and $c_{ij} \neq 0$ with probability 1.0. In Algorithm 1, after simply projecting A onto Ω, we could have

$$y_j = \sigma_1 c_{1j} v_1 + \sigma_2 c_{2j} v_2 + \cdots + \sigma_n c_{nj} v_n.$$

where σ_i is the ith singular value of A sorted by non-decreasing order such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ and $\sigma_i c_{ij}$ constitutes the weight of y_j on v_i. Consequently, random sampling ensures that all singular vectors are kept in the subspace but the singular vectors corresponding to bigger singular values likely yield bigger weights in y_j. Therefore, compared to ω_j, weights of dominating right singular vectors are amplified by the corresponding singular values. As a result, the space spanned by the columns of Y reflects dominating weights in high probability on the singular vectors corresponding to the top k singular values.

For stability consideration, Y is augmented to $m * (k + p)$ instead of $m * k$ to incorporate additional p dimensional subspace. Correspondingly, B is a $(k + p) * n$ matrix. When the SVD decomposition on B is carried out to approximate the top k singular values/vectors of A, this additional $p-$dimension space can serve as a noise-filter to get rid of unwanted subspace corresponding to small singular values. In practice, p is given with small value, such as 5 or 10, as suggested by Halko et al. [10-11].

The main computational operations in the randomized SVD algorithm involve matrix-matrix multiplications, QR decompositions, and SVD on small matrices, which are naturally fit to parallel computing platforms, for instance, distributed memory, multi-core, multi-general purpose graphics process units (GPGPU) and the Cloud computing infrastructure[11, 14].

3. Implementation

3.1 GPU-accelerated Implementation

The randomized SVD involves the following five primary computational components described in the section 2:

(1) generation of random matrix Ω;
(2) matrix-matrix multiplication of $A\Omega$ to produce Y;
(3) QR decomposition on Y;
(4) matrix-matrix multiplication of $Q^T A$; and
(5) deterministic SVD decomposition on B.

Figure 1 shows a hypothetical description of randomized SVD in finding approximate right-hand-side top- k singular vectors. The overall performance of randomized SVD depends on the efficiency of matrix-matrix multiplication, QR factorization, and SVD on small

matrices. Fortunately, after random matrix sampling by Ω, the large matrix A is condensed into either "tall-and-skinny" or "short-and-wide" matrix, such as Y and Q are $m * (k + p)$ "Tall-and-skinny" matrices, B is an $(k + p) * n$ "short-and-wide" matrix where $k + p$ is much smaller than $min(m, n)$. These small and dense matrices are particularly suitable fit in GPU memory to take advantage of high-performance computation provided. We implemented randomized SVD on GPU using CUBLAS [12] and CULA [13], and its corresponding CPU version using the Intel multi-thread MKL (Math Kernel Library) for the sake of performance illustration.

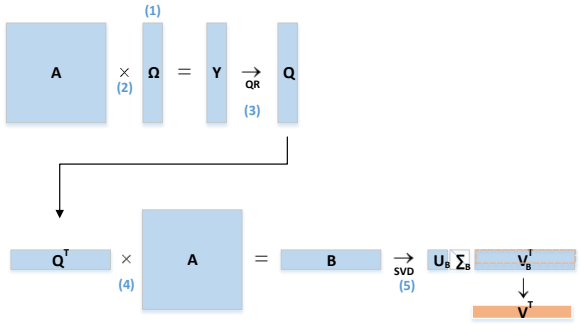


Figure 1. Procedure of Randomized SVD to Approximate Right-singular Vectors

The elapsed time spent on each primary computational component in randomized SVD is shown in Figure 2 for a 4,096 × 4,096 random matrix where k is 128 and p is 3. Multiplication between A and a "tall-and-skinny" or "short-and-wide" matrix can be efficiently carried out on the GPU's SIMT architecture and hence the computational time in generating matrix Ω and performing matrix-matrix multiplications shrinks to nearly negligible. Nevertheless, deterministic SVD, particularly when the target matrix is small, has difficulty in fully taking advantage of GPU architecture, due to a series of sequential Householder transformations need to be applied. As a result, deterministic SVD becomes the main bottleneck and thus this GPU implementation has only 1.65 over that of the CPU.

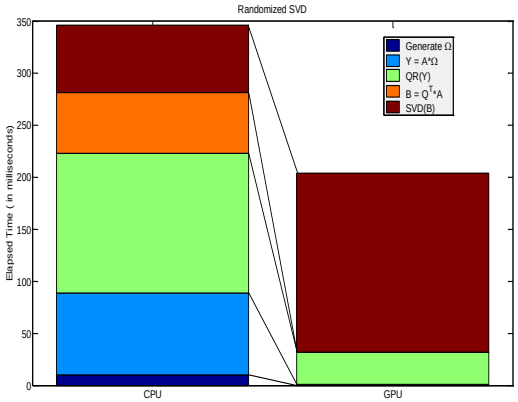


Figure 2. The Elapsed Computational Time Used in Randomized SVD on CPU-Only and GPU-Only

3.2 Approximate SVD decomposition of B

To reduce the computational cost of deterministic SVD in GPU randomized SVD implementation, we alternatively calculate the top-k singular vectors of BB^T instead of directly carrying out deterministic SVD on the "short-and-wide" matrix B. Figure 3 depicts the procedure of obtaining approximate SVD decomposition of B. Note that SVD decomposition of B is defined as

$$B = U_B \Sigma_B V_B^T.$$

Since BB^T is a small square matrix whose size is independent of the size of the original matrix A, and has SVD format as,

$$BB^T = U_B \Sigma_B U_B^T,$$

U_B could be very efficiently derived from BB^T rather than from B.

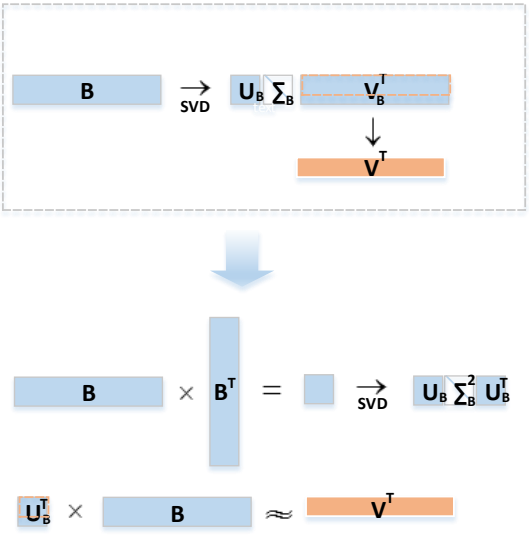


Figure 3. Procedure of Obtaining Approximate SVD Decomposition of B

Once the left singular vectors U_B become available, under the assumption that $U_B^T U_B \approx I$, where I is an identity matrix, the top k singular components could be approximated effectively through a single matrix-matrix operation

$$U_B^T B \approx \Sigma_B V_B^T.$$

Figure 4 shows the elapsed time of the improved implementation by using BB^T on the same 4,096 × 4,096 random matrix used in Figure 2. One can find that the portion of SVD computation time is significantly reduced on both CPU and GPU implementations. Consequently, the achieved speedup of GPU implementation grows up to 4.6.

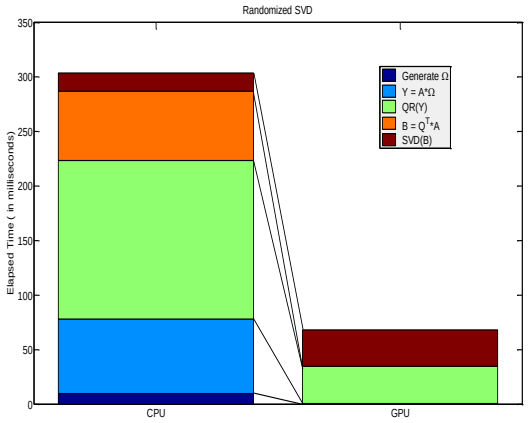


Figure 4. The Elapsed Time Used in Randomized SVD on CPU-Only and GPU-Only by Using BB^T

3.3 Hybrid GPU-CPU Scheme

As shown in figure 4, even though the alternative approach of approximating top-k singular values/singular vectors on BB^T is used, the computational time of deterministic SVD on GPU is still more than that of the CPU version due to hidden setup on GPU. To further understand the performance of deterministic SVD on GPU, we compute deterministic SVD to a set of square matrices varying in size. Figure 5 compares the computational time of deterministic SVD on CPU and GPU. One can find that the CPU implementation outperforms the GPU one on small matrices less than $2{,}500 * 2{,}500$. Therefore, using GPU to run SVD operations on small matrices is not appropriate, particularly for applications where the singular values decay very quickly and k is typically set with very small value.

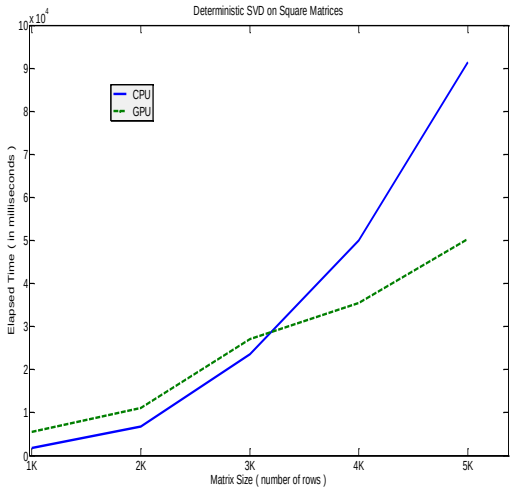


Figure 5. Comparison of Running Time for Performing Deterministic SVD on GPU and CPU

In our implementation, we develop a simple hybrid GPU-CPU scheme. If the $k * k$ square matrix is small, it will be transferred to the CPU to carry out deterministic SVD decomposition instead.

4. Results

In this section, we present the numerical results obtained with GPU-accelerated implementation on large random matrices and Mars image. The experiments are carried out on a Linux computer with an Intel Core i5-2500K CPU 3.30GHz CPU, 8GB of RAM and an NVIDIA GK110GL GPU.

4.1 Random Matrices

We generate a series of large random dense matrices of varying sizes to benchmark the performance achieved by using our GPU-accelerated randomized SVD algorithm. Figure 6 compares the computational time in logarithmic scale of performing complete SVD and randomized SVD on CPU as well as GPU-accelerated randomized SVD algorithm. The same k and p ($k = 256$ and $p = 3$) values are used. Compared to doing the complete SVD calculation on the matrix, randomized SVD has a clear computational advantage when only the top-k approximated singular components are needed. When the GPU architecture is taken advantage of, a more aggressive speedup is achieved.

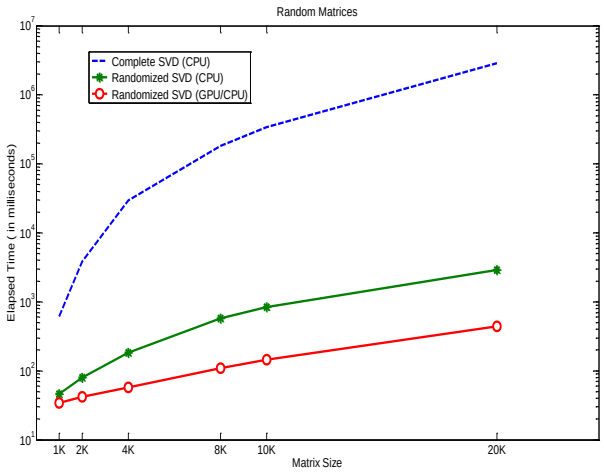


Figure 6. Comparison of Elapsed Time (logarithmic scale) of Deterministic SVD, CPU versions of Randomized SVD and GPU-accelerated Randomized SVD

Figure 7 illustrates the speedup factor for our GPU-accelerated implementation of randomized SVD over the corresponding CPU-based one. Similar to many other GPU-based algorithms, our GPU randomized SVD implementation favors larger matrices. For a $20{,}000 * 20{,}000$ matrix, the speedup can reach up to 6~7.

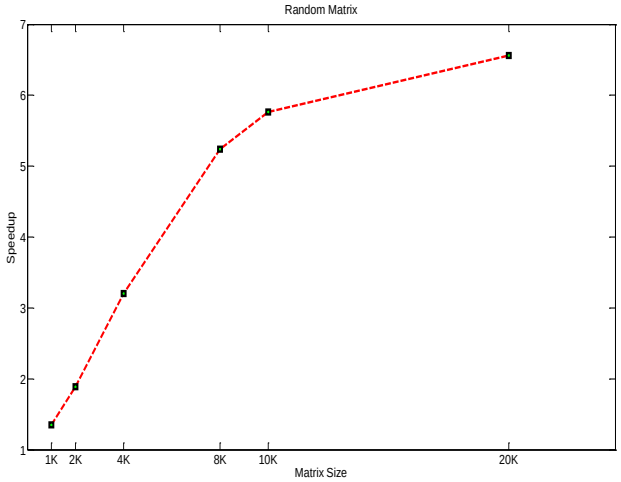


Figure 7. The Speedup of GPU-accelerated Implementation over the CPU-only Implementation.

4.2 Image Compression

We apply the randomized SVD algorithm for lossy data compression to a NASA synthesis image from the Mars Exploration Rover mission [17] shown in Figure 8. The image is an RGB $7671 * 7680 * 3$ matrix, which requires 176.74 million bytes for memory storage.

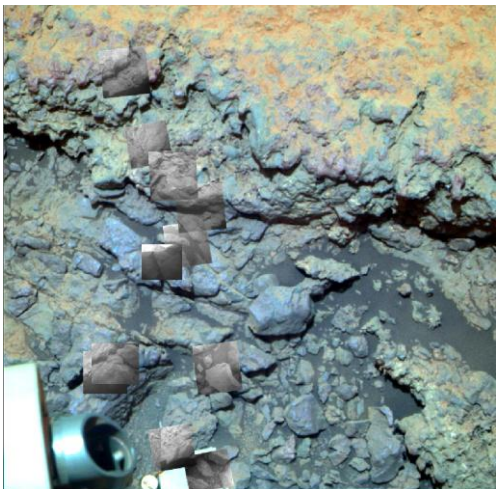


Figure 8. The Original Image

In order to compress the image, we use our GPU-accelerated implementation to obtain its low rank approximation A_k with rank 470,

$$A_k = M * N$$

where M is a 7671×470 matrix and N is a $470 * 7680$ matrix on each color channel (R,G,B). Figure 9 shows the reconstructed image, where M is computed by combining the 470 left singular vectors with the corresponding singular values while N is stored as the 470 right singular vectors as columns. To outline the effectiveness of our implementation of randomized SVD, Table 1 lists the elapsed computational time and error used in compression with Mars Image. As one can find, compared to deterministic SVD which consumes more than one thousand seconds to obtain the top 470 approximation, the GPU-accelerated randomized SVD only takes slightly more than one second. The overall storage of the decomposed image requires less than 1/8 of that of the original matrix with an acceptable 1.63% error.

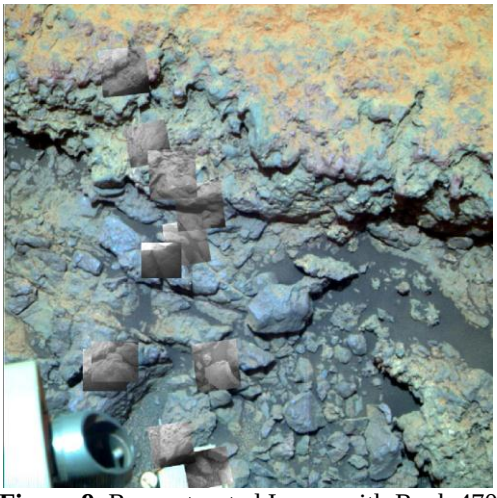


Figure 9. Reconstructed Image with Rank 470

	Elapsed Time (in seconds)	Error in Compression
Deterministic SVD	1144.71	1%
Randomized SVD	1.29	1.63%

Table 1. Elapsed Computational Time and Error in Compression with the Mars Image

5. Conclusions

In this paper, we present a GPU-accelerated implementation of randomized SVD to accelerate the process of approximating dominating singular components using both GPU and CPU. The efficiency is further improved by performing SVD decomposition on a small square matrix, which is the product of a "tall-and-skinny" matrix and its transpose. Our computational results on large random matrices and a NASA synthesis image show that the dominating singular components can be effectively obtained and the GPU-accelerated implementation outperforms the corresponding CPU version by around 6~7 times.

Acknowledgements

Yaohang Li acknowledges support from ODU 2013 Multidisciplinary Seed grant. Hao Ji acknowledges support from ODU Modeling and Simulation Fellowship.

References

[1] Alter, Orly, Patrick O. Brown, and David Botstein. 2000. "Singular value decomposition for genome-wide expression data processing and modeling." Proceedings of the National Academy of Sciences 97, no. 18: 10101-10106.

[2] Wallace, John M., Catherine Smith, and Christopher S. Bretherton. 1992. "Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies." Journal of climate 5, no. 6: 561-576.

[3] Hoecker, Andreas, and Vakhtang Kartvelishvili. 1996. "SVD approach to data unfolding." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 372, no. 3: 469-481.

[4] Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl. 2002. "Incremental singular value decomposition algorithms for highly scalable recommender systems." In Fifth International Conference on Computer and Information Science: 27-28.

[5] Calamante, Fernando, David G. Gadian, and Alan Connelly. 2000. "Delay and dispersion effects in dynamic susceptibility contrast MRI: simulations using singular value decomposition." Magnetic resonance in medicine 44, no. 3: 466-473.

[6] Golub, Gene H., and Charles F. 2012. Van Loan. Matrix computations. Vol. 3. JHU Press.

[7] Williams, Christopher, and Matthias Seeger. 2001. "Using the Nyström method to speed up kernel machines." In Advances in Neural Information Processing Systems 13.

[8] Frieze, Alan, Ravi Kannan, and Santosh Vempala. 2004. "Fast Monte-Carlo algorithms for finding low-rank approximations." Journal of the ACM (JACM) 51, no. 6: 1025-1041.

[9] Drineas, Petros, Ravi Kannan, and Michael W. Mahoney. 2006. "Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix."SIAM Journal on Computing 36, no. 1: 158-183.

[10] Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. 2011. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." SIAM review 53, no. 2: 217-288.

[11] Halko, Nathan P. 2012. "Randomized methods for computing low-rank approximations of matrices." PhD diss., University of Colorado.

[12] Nvidia, C. U. D. A. 2008. "Cublas library." NVIDIA Corporation, Santa Clara, California15.

[13] Humphrey, John R., Daniel K. Price, Kyle E. Spagnoli, Aaron L. Paolini, and Eric J. Kelmelis. 2010. "CULA: hybrid GPU accelerated linear algebra routines." InSPIE Defense, Security, and Sensing, pp. 770502-770502. International Society for Optics and Photonics.

[14] Mahoney, Michael W. 2011. "Randomized algorithms for matrices and data." arXiv preprint arXiv:1104.5557.

[15] Holmes, Michael P., Alexander G. Gray, and Charles Lee Isbell Jr. 2008. "QUIC-SVD: Fast SVD Using Cosine Trees." In NIPS, pp. 673-680.

[16] Foster, Blake, Sridhar Mahadevan, and Rui Wang. 2012. "A GPU-based approximate SVD algorithm." In Parallel Processing and Applied Mathematics, pp. 569-578. Springer Berlin Heidelberg.

[17] http://photojournal.jpl.nasa.gov/catalog/PIA14745.

Biographies

Hao Ji is a Ph.D. student in the Department of Computer Science at Old Dominion University. He received the B.S. degree in Applied Mathematics and M.S. degree in Computer Science from Hefei University of Technology in 2007 and 2010, respectively. His research interest is large-scale scientific computing.

Yaohang Li is an Associate Professor in Computer Science at Old Dominion University. He received his B.S. in Computer Science from South China University of Technology in 1997 and M.S. and Ph.D. degrees from the Department of Computer Science, Florida State University in 2000 and 2003, respectively. After graduation, he worked as a research associate in the Computer Science and Mathematics Division at Oak Ridge National Laboratory, TN. His research interest is in Computational Biology, Monte Carlo Methods, and High Performance Computing.

Intrinsically Disorder Protein Prediction using Undersampling Feedforward Neural Networks and Predicted Amino Acid Features

Qiaoyi Li[1], Steven Pascal[2], and Yaohang Li[3]

[1]Ocean Lakes High School Math and Science Academy, Virginia Beach VA
[2]Department of Chemistry, Old Dominion University, Norfolk VA
[3]Department of Computer Science, Old Dominion University, Norfolk VA

***Abstract*: In recent years, intrinsically disorder proteins have been found to be related to different types of diseases and serve important biological functions in organisms. Thus, disorder prediction has become ever more important in the understanding and modeling of these proteins. This disorder region prediction model is a neural network using around one million data samples with 525 context-based features. One problem in the use of neural network prediction methods is an unbalance of data between outputs (much lower ratio of disorder). Such data results in high discrepancies between sensitivity and specificity in order and disorder classes. By utilizing undersampling techniques to train a neural network, both sensitivity and specificity of predictions in disorder class are improved. Neural networks trained with original data reach only 25.0% sensitivity in disorder class while the ones trained with 1:1 ratio undersampling achieve 70.6% sensitivity and 76.4% specificity. Our predictions on Spinach Thylakoid Soluble Phosphoprotein (TSP9) and Prostate apoptosis response factor-4 (Par-4) agree with the NMR experimental results.**

Keywords: Intrinsically disorder protein, prediction, neural network, balanced training set

1. INTRODUCTION

A challenge in protein structure prediction is the recognition and modeling of intrinsically disorder proteins (IDPs), also called intrinsically unstructured or naturally unfolded proteins. Disorder regions of proteins do not display a stable tertiary structure when the polypeptide is isolated in vitro. Their dynamic structure results from frequent variation of phi and psi angles in the protein backbone. Due to the unique, flexible nature of intrinsically disorder proteins, they can serve important biological functions and play a role in neurological disease.

Theoretically, IDPs contain a high proportion of polar and charged amino acid residues, which prevent the protein from establishing a stable globular structure. Disorder proteins are often low-complexity, featuring repetition of a few amino acids. However, not all low-complexity protein sequences are intrinsically disorder. Another characteristic of IDPs is low contents of stable secondary structure such as α-helices and β-sheets. All disorder regions are found within segments of coiled secondary structures (loops). Some of these structures are called "hot loops" [Linding et al., 2003] due to their mobility from high C-α temperature factors. While structured proteins do move continuously due to kinetic and thermal energy, disorder polypeptides have much higher B-factors signifying a more dynamic structure.

The greatest advantage of IDPs is their flexibility, which facilitates binding to target molecules of various sizes and shapes. Intrinsic disorder occurs more frequently in proteins involved in cell signaling, transcription, and chromatin remodeling. Some disorder proteins link together two globular or trans-membrane domains. IDPs also form fuzzy complexes where their structural uncertainty is essential for function. In coupled folding and binding, disorder proteins fold into a more stable structure after attachment to another macromolecule. The combination of folding and binding enhance selective abilities of the molecules and speed up the binding process.

There are many existing methods for the prediction of disorder regions in proteins. Most methods of prediction mostly consider amino acid sequencing information (*ab initio* methods). An early ab initio method SEG used areas composed of low complexity of sequencing with limited success since not all low complexity sequences are disorder. Some early sequence based predictors (e.g. HCA, PreLink, FoldIndex) use mainly hydrophobicity and hydropathy to predict disorder based on their low content of hydrophobicity. However, this method is more suited to predicting shorter segments of disorder [Deng et al., 2012]. GlobPlot takes a simple approach by using a running sum of the amino acid's propensity to be disorder or ordered (from Russell/Linding scale of disorder) to predict the segment's propensity for globularity. The propensities are then plotted to show different areas of disorder/order. The method's simplicity only allows it to provide a general estimate [Linding et al., 2003]. PONDR, a collection of three predictors (VL-XT, XLT, and CaN) incorporated and

combined other features of disorder such as flexibility as well as sequence complexity and hydropathy in the analysis of amino acid sequencing through a feedforward neural network [Li et al., 1999]. DisEMBL uses a neural network trained on structure data from X-ray crystallography that can predict regions of loops and "hot loops" with 64% sensitivity [Linding et al., 2003]. The PSI-BLAST algorithm has contributed to the improvement of predictions. Based on CASP10 results in 2012, Diopred3 was ranked second in Disorder Protein Prediction. It uses PSI-BLAST to create sequence profiles for each protein target which then trains linear support vector machines, reaching accuracies of 70%. Prdos-CNF [Eickholt and Cheng, 2013], the highest ranked overall disorder protein predictor based on the CASP10 report, uses a recursive neural network with inputs of PSI-BLAST arranged profile, predicted secondary structure, and solvent accessibility. CASP10 results showed an accuracy of 71% and a precision of 70%.

Other predictor methods include clustering and meta methods. Clustering methods including DISOclust [McGuffin, 2008] predict a tertiary structure which is then layered over the target protein to calculate a probability of disorder. These meta methods include metaPrDOS2, which acquired the best accuracy of 77.8% in CASP10 with a sensitivity of 64.73% and a specificity of 89.4% [Monastyrskyy et al., 2014].

2. METHODS

2.1. Dataset

Overall, 121,298 disorder residue samples are obtained from the disorder protein database Disprot [Sikmeier et al., 2007]. 1,010,750 structured residue samples are extracted from a set of chains with maximum 25% pair-wise sequence identity, 1.8A resolution, and 0.25 R-factor generated by the PISCES server [Wang and Dunbrack, 2003]. These data samples are mixed together and then are randomly split into two disjoint sets. Using a preprogrammed feedforward neural network training application in MATLAB written for pattern recognition, the first fold and second fold data sets are applied alternatively for training and validating the neural network of 200 neurons (i.e. first fold is used in training and second fold for testing and vice versa).

2.2. Neural Network Encoding

For each data sample, a sliding window of 21 residues is selected, where the feedforward neural network is trained to predict if the centered residue in the window is disorder. Each residue is represented by 20 PSSM values, 1 boundary value indicating C- or N-terminals overlap, 3 values of secondary structure probabilities predicted by SCORPION [Ashraf and Li, 2014], and 1 value of solvent accessibility probability predicted by CASA [Ashraf and Li, 2014]. Putting every feature together, there are totally 525 features associated with each residue sample. Figure 1 shows the neural network encoding scheme.

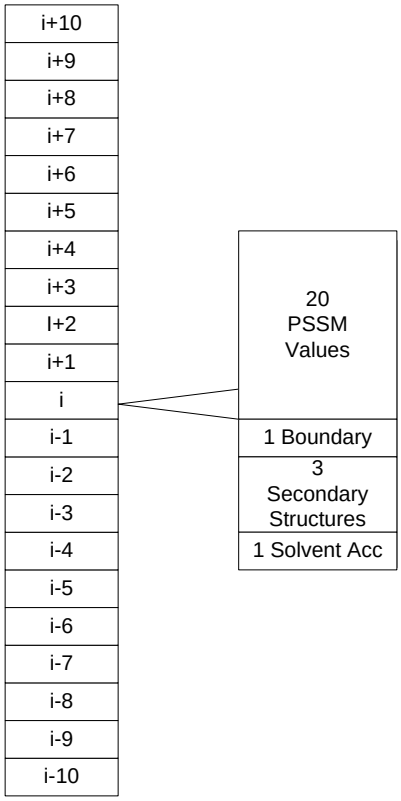


Figure 1. Neural Network Encoding Scheme for Intrinsically Disorder Region Prediction

2.3. Balancing the Training Set

Due to proportionally low occurrence of IDPs, the training was unbalanced. A low ratio of one output class results in higher identification accuracy for the majority output and lower accuracy of the minority output. The method undersampling alleviates unbalanced training by utilizing fewer structured data samples in the training data set. Undersampling ratios of ordered to disorder data samples used to train the neural networks are 1:1, 3:1, and 6:1 respectively. All three undersampling ratios are applied to both data sets yielding six unique sets. Each set is used to train the feedforward neural network and non-undersampled sets are utilized for cross validation.

3. RESULTS

3.1. Prediction Accuracy and Sensitivity

Table 1 compares sensitivity and specificity of the trained neural networks using various undersampling strategies. In the original training data, the neural network

predictions biases to the class of "ordered" residue because the ordered residue samples significantly outnumber the disorder ones. When the undersampling strategies are employed, such sampling biases are reduced. Hence both sensitivity and specificity of disorder residue predictions increase with the price of relatively small reduction of sensitivity and specificity of ordered residue predictions. When the numbers of ordered and disorder samples are approximately equal (1:1 undersampling), the sensitivity and specificity of both disorder and ordered predictions are above 70%.

Table 1. Comparison of original and undersampling methods in neural network training and testing

		Sensitivity (%)		Specificity (%)		Total Acc.
		Disorder	Ordered	Disorder	Ordered	
Original	Train	26.0	98.7	68.9	92.3	91.4
	Test	25.3	98.6	66.8	92.3	91.3
Undersampling 6:1	Train	32.3	98.0	72.8	89.6	88.5
	Test	31.7	97.9	70.8	89.9	88.7
Undersampling 3:1	Train	46.3	94.5	73.8	84.0	82.4
	Test	45.2	94.3	72.4	83.9	82.1
Undersampling 1:1	Train	71.0	78.3	76.6	73.0	74.7
	Test	70.6	77.8	76.4	72.3	74.2

3.2. Prediction on Spinach Thylakoid Soluble Phosphoprotein (TSP9)

Thylakoid Soluble Phosphoprotein (TSP9) (PDB ID: 2FFT) is a plant-specific protein in the photosynthetic thylakoid membrane, which is disorder under aqueous conditions discovered by NMR spectroscopy [Song et al., 2006]. Figure 2 depicts a 3D NMR model of TSP9. The 15-23 residues form a small α-helix, which is a potential binding site and the rest of the protein chain is disorder. Figure 3 shows the secondary structure assignment by NMR and the disorder regions predicted by our neural network trained with 1:1 undersampling. One can find that our neural network is able to correctly identify the region of the small α-helix as ordered and the majority of the disorder loops. There are a few mispredictions in the disorder loops, which may be corrected by an additional trained neural network for refinement.

3.1. Prediction on Prostate apoptosis response factor-4 (Par-4)

Prostate apoptosis response factor-4 (Par-4) is a proapoptotic protein coding for tumor-suppression. We apply our trained neural network for disorder region prediction with 1:1 undersampling to Par-4, which is shown in Figure 4. Our neural network predicts that residues from 31 to 138 belong to a long disorder region, which agrees with the experimental results [Libich et al., 2009] observed by NMR spectroscopy.

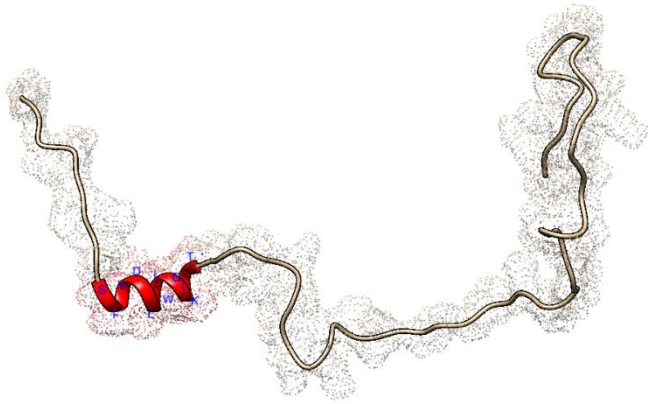


Figure 2. 3D NMR Model of Spinach Thylakoid Soluble Phosphoprotein (TSP9)

```
1     SAAKGTAETK QEKSFVDWLL GKITKEDQFY
                    SHHHHHH HHHS    S
      DDDDDDDDDO OOOOOOOOOO OOODDDDDDO

31    ETDPILRGGD VKSSGSTSGK KGGTTSGKKG
       S  SSS S      S         S S S  S
      DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

61    TVSIPSKKKN GNGGVFGGLF AKKD
      SSSS SS S   S  SSSSSS
      ODDDDDDDDD DDDDOOOOOO ODDD
```

Figure 3. Secondary Structure Assignment and Predicted Disorder Regions in Spinach Thylakoid Soluble Phosphoprotein (TSP9)

```
1     MATGGYRSSG STTDFLEEWK AKREKMRAKQ
      DDDDDODDDD DDOOOOOOOO OOOOOOOOOO

31    NPVGPGSSGG DPAAKSPAGP LAQTTAAGTS
      DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

61    ELNHGPAGAA APAAPGPGAL NCAHGSSALP
      DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

91    RGAPGSRRPE DECPIAAGAA GAPASRGDEE
      DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

121   EPDSAPEKGR SSGPSARKGK GQIEKRKLRE
      DDDDDDDDDD DDDDDDDOO OOOOOOOOOO

151   KRRSTGVVNI PAAECLDEYE DDEAGQKERK
      OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO
```

```
181   REDAITQQNT IQNEAASLPD PGTSYLPQDP
      OOOOOOOOOO OOOOOOOODD DOOOOODDOO

211   SRTVPGRYKS TISAPEEEIL NRYPRTDRSG
      OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

241   FSRHNRDTSA PANFASSSTL EKRIEDLEKE
      OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

271   VLRERQENLR LTRLMQDKEE MIGKLKEEID
      OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

301   LLNRDLDDME DENEQLKQEN KTLLKVVGQL
      OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

331   TR
      OO
```

Figure 4. Disorder Region Prediction in Par-4

4. CONCLUSIONS

Using a pre-established, unspecialized MATLAB pattern recognition application for neural network training, the network has achieved a best sensitivity of 70.6% and a best total accuracy of 76.4% in predicting disorder residues. Best results originated from using 1:1 ratio under sampled data. By balancing neural network training through undersampling, sensitivity and specificity became closer to each other thus improving the sensitivity where networks would usually be skewed toward structured proteins. Our predictions show good agreements with the NMR experimental results on Spinach Thylakoid Soluble Phosphoprotein (TSP9) and Prostate apoptosis response factor-4 (Par-4).

There is a lot of space for this method to be improved. For example, using oversampling instead of undersampling on a powerful computer may lead to further prediction accuracy improvements. Incorporation of additional predicted features such as B-factor profiles, inter-residue contacts, and disulfide bonding states may be helpful for training a more effective neural network. Advanced machine learning algorithms such as deep learning, if properly used, may also lead to advancement in disorder region prediction. All these will be our future research directions.

REFERENCES

1. B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, K. Fidelis, "Assessment of Protein Disorder Region Predictions in CASP 10." *Proteins: Structure, Function and Bioinformatics*, 82: 127-137, 2014.
2. X. Li, P. Romero, M. Rani, A. K. Dunker, Z. Obradovic, "Predicting Protein Disorder for n-, c-, and Internal Regions," Genome informatics, 10: 30-40, 1999.
3. R. Linding, L. J. Jenson, F. Diella, P. Bork, T. J. Gibson, R. Russell, "Protein Disorder Prediction: Implications for Structural Promteomics," Structure, 11: 1453-1459, 2003.
4. R. Linding, R. Russell, V. Neduva, T. Gibson, "Globplot: Exploring Protein Sequences for Globularity and Disorder," Nucleic Acid Research, 31: 3701-3708, 2003.
5. X. Deng, J. Eickholt, J. Cheng, "A Comprhensive Overview of Computational Protein Disorder Prediction Methods," Mol. Biosyst., 1: 114-121, 2012.
6. A. Yaseen, Y. Li, "Context-based Features Enhance Protein Secondary Structure Prediction Accuracy," Journal of Chemical Information and Modeling, in press, 2014.
7. A. Yaseen, Y. Li, "CASA: A Protein Solvent Accessibility Prediction Server using Context-based Features to Enhance Prediction Accuracy," BMC Bioinformatics, in press, 2014.
8. J. Song, M. S. Lee, I. Carlberg, A. V. Vener, J. L. Markley, "Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications," Biochemistry, 45: 15633-15643, 2006.
9. D. S. Libich, M. Schwalbe, S. Kate, H. Venugopal, J. K. Claridge, P. J. B. Edwards, K. Dutta, S. M. Pascal, "Intrinsic disorder and coiled-coil formation in prostate apoptosis response factor 4," the FEBS Journal, 276: 3710-3728, 2009.
10. M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, A. K. Dunker, "DisProt: the Database of Disorder Proteins," Nucleic Acids Res., 35: D786-793, 2007.
11. G. L. Wang, R. L. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 19: 1589-1591, 2003.
12. J. Eickholt, J. Cheng, "DNdisorder: predicting protein disorder using boosting and deep networks," BMC Bioinformatics, 14: 88, 2013.
13. L. J. McGuffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models," Bioinformatics, 24: 1798-804, 2008.

Thread Affinity, Power, Energy, and Performance on the Intel Xeon Phi

Gary Lawson and Masha Sosonkina, Ph. D
Department of Modeling, Simulation, and Visualization Engineering
Old Dominion University
1300 Engineering and Computer Science Bldg
Norfolk, VA 23529
glaws003@odu.edu, msosonki@odu.edu

Keywords: Intel Xeon Phi, thread affinity, energy efficiency, EP benchmark

Abstract

This work investigates the impact of varying thread affinity on performance, power and energy consumption for the Intel Xeon Phi. To stress the Xeon Phi, the embarrassingly parallel application (class C) from the NASA Advanced Supercomputing (NAS) Division is natively executed. Power and performance data is collected and compiled for various thread count and affinity combinations. An increase in performance was observed as the maximum number of threads was utilized for *compact* and *balanced* affinity types throughout the experiment. The *scatter* affinity type was observed to perform well while the number of threads remained low (<= 180); however its performance quickly degrades after this "threshold".

The lowest observed run *balanced*, required 235 threads and consumed 6.66% less energy as the *None* affinity type, which is default on the Intel Xeon Phi. The execution time for the *balanced* affinity was 4.32% lower than the *None* affinity as well.

1. INTRODUCTION

The Intel Xeon Phi is a coprocessor which utilizes Intel's Many Integrated Core (MIC) architecture to speed up parallel processes. The Xeon Phi is also known as an accelerator and is composed of 60 cores at 1.05 GHz. Each core is capable of simultaneously processing 4 threads in hardware [1]. In this work, no optimization was considered for the NAS benchmark application; however it is important to note that the Xeon Phi performs best with calculation intensive applications which can take advantage of the Vector Processing Unit (VPU). The VPU features a 512-bit SIMD instruction set; 16 single-precision and 8 double-precision operations may be computed per cycle. The VPU also supports Fused Multiply-Add (FMA) instructions pushing the simultaneous operations per cycle to 32 and 16 respectively [1].

This work focuses on the impact of varying the processor affinity options on performance and power.

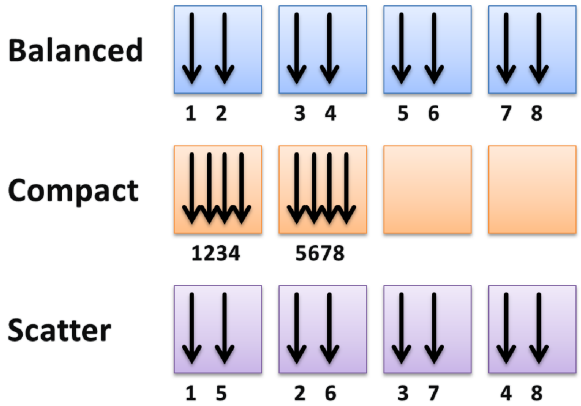


Figure 1. An example of various thread affinity policies.

Affinity is described as a set of policies which determine how threads/processes are bound to cores [2]. For OpenMP and MPI, a series of environment variables may be assigned for this purpose. OpenMP is used in this work as its parallelization strategies may be used directly by the Xeon Phi. MPI is not considered.

The Xeon Phi supports six affinities policies: *balanced*, *compact*, *scatter*, *none*, *disabled*, and *explicit* [3]. However, in this work, only *balanced*, *compact*, *scatter*, and *none* are tested. *Disabled* disables the thread affinity interface and *explicit* allows specific assignment of threads. Figure 1 provides a graphical description of how threads are distributed to processors for the various affinity types considered. The *balanced* affinity policy evenly distributes threads amongst the cores. This policy will attempt to utilize all available cores while keeping OpenMP threads close to one another. The *compact* affinity policy prioritizes filling each core with the maximum number of threads. This policy does not necessarily utilize every core in the system; it varies depending on the thread count. Finally, the *scatter* affinity distributes threads to every core in a round-robin fashion. This affinity is useful in cases where cache is not shared locally between threads [2]. Affinity type *none* is default; it allows the Xeon Phi device full control over thread grouping and distribution amongst threads and is used as the control in our experiment.

```
[glawson@borges Affinity]$ micsmc -f

mic0 (freq):
Core Frequency: .......... 1.05 GHz
Total Power: ............. 37.00 Watts
Lo Power Limit: .......... 257.00 Watts
Hi Power Limit: .......... 306.00 Watts
Phys Power Limit: ........ 326.00 Watts


mic1 (freq):
Core Frequency: .......... 1.05 GHz
Total Power: ............. 79.00 Watts
Lo Power Limit: .......... 257.00 Watts
Hi Power Limit: .......... 306.00 Watts
Phys Power Limit: ........ 326.00 Watts


[glawson@borges Affinity]$
```

Figure 2. Sample MICSMC power and frequency output information.

In addition to affinity, threads may be controlled more directly through another variable, granularity. Granularity determines at what level the policies are applied on the Xeon Phi. Granularity consists of three levels: *fine*, *thread*, or *core*. The *fine* and *thread* granularity levels are similar in that they bind threads to a single context [3, 4]. The *core* granularity binds threads to a core; however the threads may float within the context of the physical core [4]. This work uses the *thread* granularity level because it is most specific.

It has been observed that most discussions of thread affinity involve the use of matrix multiplication examples [4, 5]. In this work, the embarrassingly parallel benchmark application is used, courtesy of the NAS [6]. This work also incorporates preliminary results from the CG (Conjugate Gradient) benchmark; however further testing is left for future work. The embarrassingly parallel application generates pairs of Gaussian random deviates according to a specific scheme and attempts to establish a reference point for peak performance of a given platform [7]. The NAS EP benchmark is compiled for class C problem sizes to run natively on the Intel Xeon Phi. An application is executed natively if it was compiled specifically for the Xeon Phi device (-mmic). Additionally, the conjugate gradient application tests unstructured grid computations and communications. The matrix is provided with randomly generated locations, one per entry. Each entry is computed by an approximation to the smallest eigenvalue of the large, sparse, unstructured matrix [7].

The remainder of this paper is outlined as follows: Section 2 discusses the method used to read frequency and power data from the Xeon Phi. Section 3 will discuss the design for the affinity experiment itself and Section 4 will present the experimental results and a discussion of their impact on performance and energy. Section 5 presents an analysis of the results as well as a preliminary analysis on the CG test results. Section 6 will conclude this document.

2. POWER DATA EXTRACTION METHOD

Extracting the power data from the Xeon Phi required use of Intel's MIC System Management and Configuration (SMC) utility tool, *micsmc* [8]. One goal of this work is to determine whether or not micsmc is suitable for monitoring and performing analysis on an application in real time. Through the Linux command shell, the tool may be used to print information to the console, as displayed in figure 2.

The output from the micsmc tool is redirected to a C++ application, powerMon, which parses the output in real time and outputs only the necessary data: frequency, total power, total time, time elapsed since previous update, and mic id. This data is redirected to a file through the Linux shell. In the experiment, the micsmc elapsed time between updates was observed to range between 80 ms and 110 ms. This range supplied a sufficient time-slice when calculating total energy.

3. EXPERIMENT DESIGN

The experimental platform is a small cluster 'Borges' which contains two 8-core Intel Xeon E5-2650's, 2 GHz. The platform also contains two Intel Xeon Phi's; each with 60 cores at 1.05 GHz. The host contains 64 Gb of RAM; and each Xeon Phi contains 8 Gb of memory. This work only requires a single mic, and minimal memory and processing power on the host system. Energy consumed was *only* calculated for the Intel Xeon Phi during the execution of the benchmark test.

Table 1. Energy, power, and performance data for 236 threads vs 240 threads.

Compact Affinity			
	Thread 236	Thread 240	Difference (%)
Energy (J)	2674.69	2879.18	+ 7.10
Power (W)	131.11	139.30	+ 5.88
Performance (mops)	555.71	531.89	- 4.48
Execution Time (s)	15.46	16.15	+ 4.29

The experiment itself is simple: determine the impact thread affinity has on performance, power, and energy. It consists of running the benchmark application a few thousand times; for each affinity and for each thread count from 50 to 236 run the Affinity Test script provided the input parameters: affinity, thread count, benchmark, and class size. According to Intel [9], OS processes are only mapped to the last core, 60 in this work, and it is suggested the core remains free. Therefore, this work uses a max of 236 threads. Table 1 presents the energy, power, and

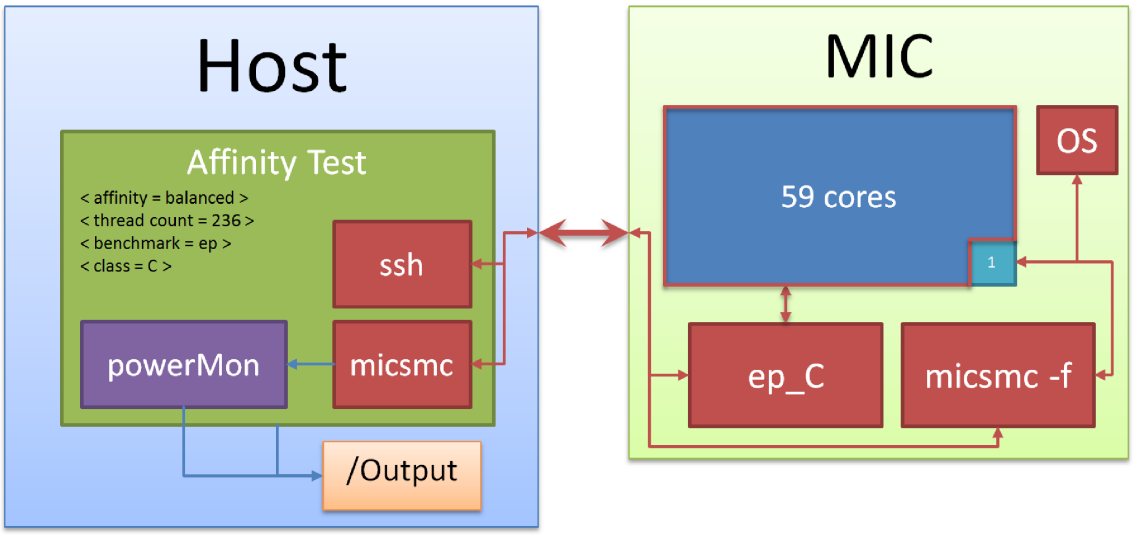


Figure 3. Experiment Overview featuring the Host's shell and the Intel Xeon Phi or 'MIC'. Red connections describe bidirectional communication while blue connections describe unidirectional communication.

performance analysis for allocating additional threads, 240 as opposed to 236. Immediately from table 1, a difference of 7.10% energy consumption increase is observed when allocating additional threads which would occupy the final core. Performance is also observed to decrease approximately 4%. To avoid these performance and energy penalties, the maximum is 236 threads.

Figure 3 provides an overview of the test itself and the interactions between the Host shell and the Intel Xeon Phi. Within the host are a set of applications and scripts which orchestrate the tests. The figure provides an example of a single iteration; the iteration details are described inside the "Affinity Test" block. In this example, the *balanced* affinity is running with 236 threads. The benchmark used throughout this work is EP (embarrassingly parallel) and its class is size C. The affinity test script spawns several applications: micsmc, powerMon, and ssh. The micsmc utility tool for reading power information on the mic is started along with the powerMon application. More details on power extraction are presented in the Section 2. The ssh connection to the mic sends a script which is run by the device. Inside the script, the environment variables are set, and then the benchmark application is run; the application was compiled offline and transferred to the device. Output which normally would appear on the console has been redirected to the /Output folder. The output from the EP_C application on the mic is stored in a single log file while the output from the powerMon application is stored in individual *power data* files.

After the experiment has completed, the power data files which store the time-slice power, frequency, and timing information are processed: minimum and maximum readings for power, time elapsed, and frequency are collected, average power and time elapsed are computed, the mode for frequency is collected, and energy is calculated. These data items are stored in CSV (Comma-Separated Value) files for analysis.

4. EXPERIMENTAL RESULTS

This section discusses the results observed from the affinity experiment. The results are broken into the following subsections: Execution Time results, Performance results, Average Power results, and Total Energy Consumption results.

4.1. Execution Time

Figure 4 presents the graph detailing the execution timing results. By observation, it is immediately obvious that *compact* is extremely inefficient at low thread counts. This may be attributed to the application itself; the embarrassingly parallel benchmark is computationally intensive. For thread counts less than or equal to 180, thread contention may be the culprit behind *compact*'s poor performance. However, after this threshold, *scatter* becomes less efficient and *balanced*/*compact* are practically identical. This is to be expected; for *scatter*, threads are spread far apart from one another (see Figure 1). It may be that *scatter* becomes overwhelmed by caching overhead after 180 threads. *Balanced* and *compact* become more identical as the thread count increases with relation to thread mappings. At 236 threads, the thread mappings for the *balanced* and *compact* affinities are identical. *None* seems to follow *balanced*, however its execution is not as stable. This instability is present in all of the results.

Figure 4. Execution time (seconds) for the various affinities.

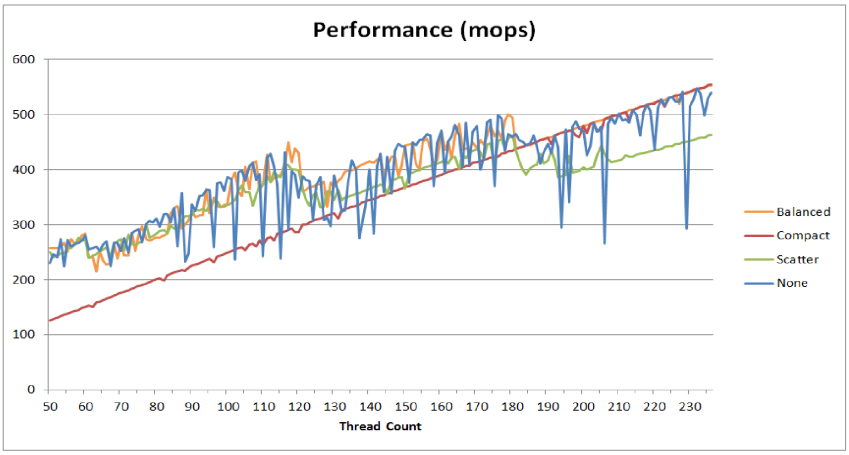


Figure 5. Performance (mops) for the various affinities.

4.2. Performance

In figure 5, the performance measurements for the embarrassingly parallel benchmark are compared. The figure contains the performance measured in mops (million operations per second); the graph displays the total mops for the benchmark execution for the various affinities and thread counts. Again after 180 threads, the *scatter* affinity plummets; this time in performance. Also, interestingly, the compact affinity is observed to have a consistent, linear increase as thread count increases; this is ideal. Comparing figures 4, 5, and 6 with respect to *none* and *balanced*, a pattern is observed. It appears the *none* affinity mimics the performance of the *balanced* affinity; however it is severely unstable. This instability may be due to the lack of rules governing the *none* affinity.

4.3. Average Power

Figure 6 displays the computed average power in Watts. The average power is calculated by averaging over the power readings stored in the *power data* output file. Observing figure 6, the range for average power is interesting. A maximum of 139 Watts has been achieved; however the Xeon Phi supports wattages of 300+ (see figure 1). Further research would need to stress the system further with a longer duration test, and a more demanding problem such that higher power ratings may be observed. *Compact* is observed with a fairly linear increase, and *scatter* is observed to plummet after 180 threads. However, a lower average power is a positive result, and this is a promising result for the *scatter* affinity. Although the performance has been lower, so has the power; figure 7 presents the details on total energy consumed while running the EP_C benchmark application natively on the Xeon Phi. In figure 7, it will be determined whether or not *scatter* is viable affinity for this type of application.

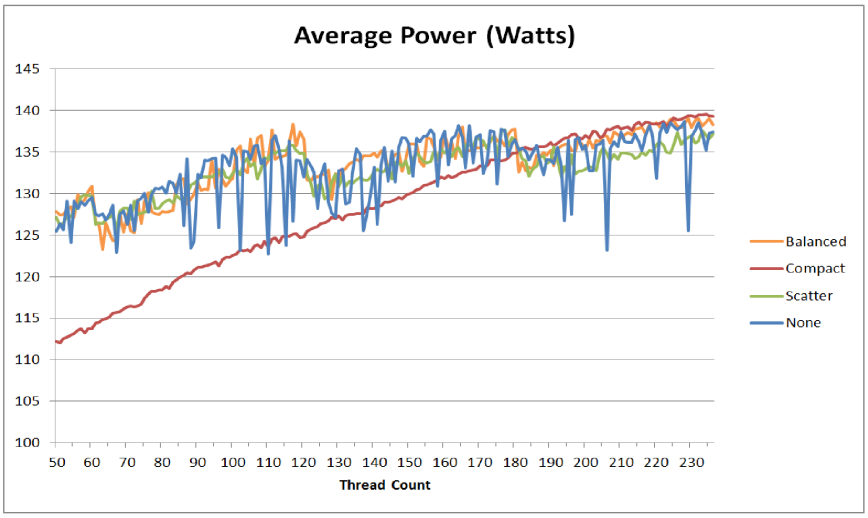


Figure 6. Average power (Watts) for the various affinities.

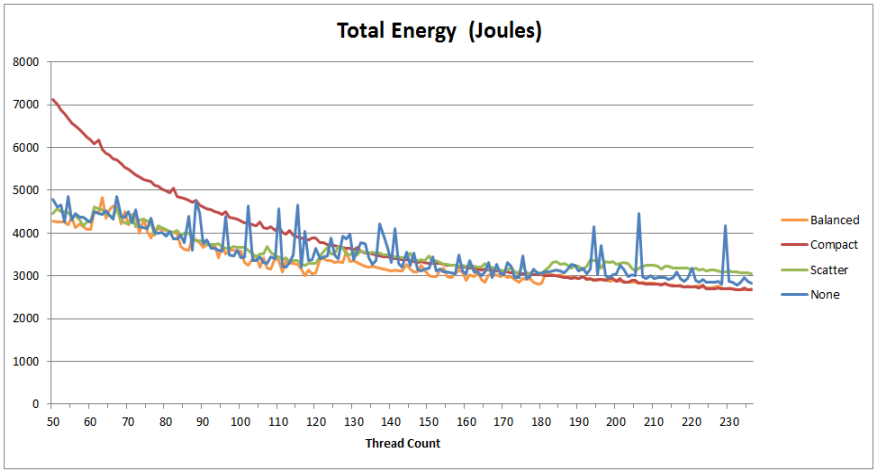


Figure 7. Total energy (Joules) for the various affinities.

4.4. Energy

In figure 7, the *compact* affinity is an extreme outlier for fewer than 180 threads. However, after this threshold, the *compact* and *balanced* affinities are observed to provide the most efficient solution. *None*, the default affinity, performs fairly well and is followed by *scatter*. The most efficient solution is the *balanced* affinity with 235 threads which had an energy cost of 2673.1 J, as shown in table 2. The second most efficient solution is detailed in table 1: *compact* affinity with 236 threads at 2674.7 J. Table 2 presents an analysis comparing the lowest consumed energy measurement and the results for the same thread count for *none*. Ultimately, *scatter* is an affinity best utilized with fewer threads per application, and more applications per Xeon Phi.

5. ANALYSIS

This section presents the analysis for the results from Section 4. It also presents preliminary data and a short results and analysis on the data.

5.1. Thread Affinity Test Analysis

The results in table 2 are promising; Balanced 235 is observed 6.66% lower in energy consumption, with a performance increase of 4.32% in execution time, and a performance decrease of 4.51% in mops. Average power was higher by 1.32% compared to None 235's average. These results show that optimizing affinity can provide significant energy and performance increases for the EP_C benchmark. This result could be applied to similar applications; total energy would be expected to decrease, as well as execution time. Recent energy efficiency research

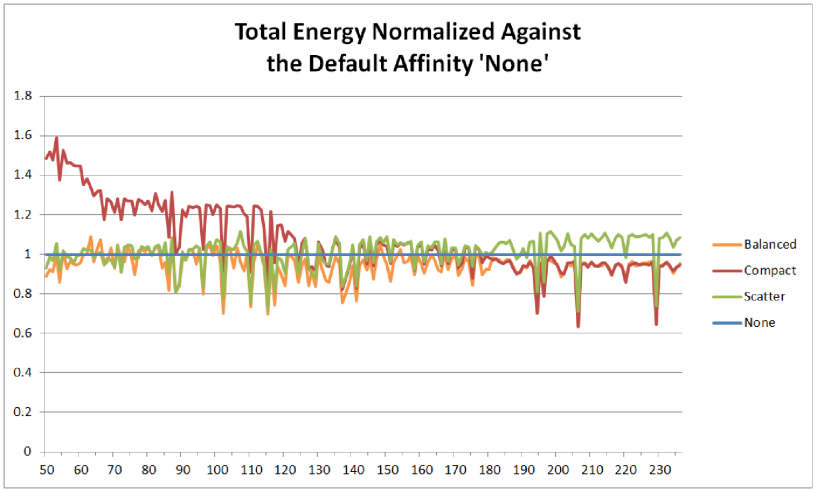


Figure 8. Total energy normalized against the default affinity type, none.

performed on the CPU record the EP benchmark experiencing slight performance loss <1% and used slightly more energy <1% while attempting to use a general energy saving strategy [10]. Although the energy saving strategy did not perform well with the EP benchmark, it performed well over all cases considered. The results in this work show that the Intel Xeon Phi can offer performance and energy gains to applications such as EP.

Figure 8 presents the normalized results comparing the *balanced*, *compact*, and *scatter* affinities to *none*. *None* is consistently 1, and is omitted from figure 8. *Compact*, at the lower thread counts, is observed to be the highest cost affinity with *balanced* and *scatter* being more energy efficient alternatives. Between 130 and 180 threads, *compact*, *balanced*, and *scatter* become approximately equivalent in that they are all positive and negative with respect to *none* for the range described. However, *balanced*, *compact*, and *scatter* stay within a range of 0.1, or 10% total energy. Finally, after 180 threads, *scatter* and *balanced*/*compact* diverge; *scatter* becoming less efficient than *none*, and *balanced*, *affinity* become more efficient than *none.* At the peak, *balanced*, *compact*, and *scatter* are 45%, and 35% (*scatter*) more efficient than *none*.

5.2. Preliminary Results & Analysis

In figure 9, energy consumption is presented comparing the EP benchmark for the class C problem size and the CG benchmark for the class B problem size. The figure displays four bars, each representing the affinity for a varying test (CG_B Balanced, CG_B Compact, EP_C Balanced, and CP_C Compact). The thread counts 59, 118, 177, 236 represent the energy levels for each test; the thread counts are equivalent to 1, 2, 3, and 4 threads per core respectively.

Interestingly, the CG_B test (blue and red) requires as much energy as the compact affinity for the EP_C test.

Table 2. Lowest consumed energy measurements vs default affinity type.

	Balanced 235	None 235	Difference (%)
Energy (J)	2673.10	2863.97	+ 6.66
Power (W)	139.04	137.23	- 1.32
Performance (mops)	553.1	529.23	- 4.51
Execution Time (s)	15.53	16.23	+ 4.31

However, neither affinity was as efficient as EP_C Balanced. This may be attributed to the communication and memory contentions. Total memory nor communication time were measured in this work, but may be pursued in the future. It is also observed that both tests require approximately the same amount of energy; and for both applications, the compact affinity provides the lowest energy consumption. This is not the minimum consumption over all data, just a minimum between the four thread count cases.

Figure 10 presents the preliminary results for the execution time between the CG class B and EP class C benchmarks. Similar to figure 9, both affinities for CG require more time to complete. However, the rate of decay between adding an additional thread per core is exponential. This phenomenon does not yet have an explanation; speculation points to memory contention or communication or a mixture of both requiring more cores and threads for both affinities.

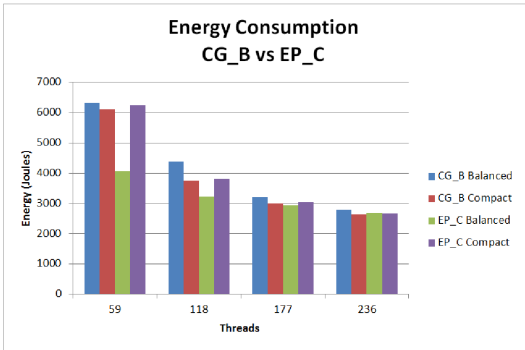


Figure 9. Preliminary results for energy consumption (Joules) between CG_B and EP_C.

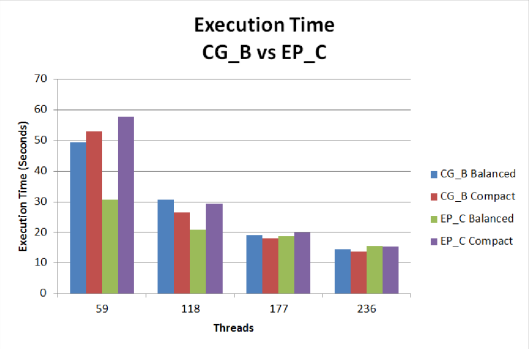


Figure 10. Preliminary results for execution time (Seconds) between CG_B and EP_C.

6. CONCLUSIONS

This work sought out to determine what impact affinity has on applications with interest in performance, measured in mops, power, measured in watts, and energy, measured in Joules. The balanced affinity at 235 threads was observed to perform the best with the lowest energy consumption measurement. When compared to the default affinity, none, balanced 235 was 6.66% lower in energy consumption and 4.32% faster in execution time. None 235 was observed to have a 1.32% lower average power and a 4.51% higher performance. These results are promising for applications similar to EP_C, the embarrassingly parallel class C benchmark. In general, leaving affinity to the Xeon Phi is not a good option. Developers can code it into their applications [3], and users can include the export option: 'export KMP_AFFINITY=granularity=thread, balanced'.

This work was also interested in determining whether or not micsmc was a viable tool for monitoring power and frequency data on the Intel Xeon Phi. It was determined that a time-slice range of 0.08 to 0.11 seconds was acceptable for providing accurate power reads such that total energy could be calculated more accurately.

Additionally, this work presented a preliminary analysis comparing the EP to the CG benchmark; both courtesy of NAS. The results suggested that communication intensive and memory intensive applications benefit from maximum utilization to obtain lower energy costs.

Future studies involve researching and possibly implementing a method for measuring power and performance directly through the MSR's (Model Specific Register). Additionally, further explanation and analysis for comparing the results for EP to other benchmarks, including CG will be explored in future studies. Research, implementation, and experimentation on vectorization optimization will be pursued and discussed in future work.

REFERENCES

[1] Chrysos, George. "Intel Xeon Phi Coprocessor – the Architecture". Intel Corporation. 2012. http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-codename-knights-corner

[2] Birkland, Aaron. "Cornell Virtual Workshop". Cornell Center for Advanced Computing. 2013. https://www.cac.cornell.edu/vw/mic/default.aspx

[3] Intel Corporation. "Intel C++ Compiler 12.1 User and Reference Guides". 2011. http://software.intel.com/sites/products/documentation/studio/composer/en-us/2011Update/compiler_c/hh_goto.htm#main/main_cover_title.htm

[4] Barth, Michaela, et. al. "Best Practice Guide Intel Xeon Phi v1.1". Prace. 2014. http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Intel Xeon-Phi.pdf

[5] Masci, F. "Benchmarking the Intel Xeon Phi Coprocessor". Infrared Processing and Analysis Center, Caltech. 2013. http://web.ipac.caltech.edu/staff/fmasci/home/miscscience/MIC_benchmarking_2013.pdf

[6] Dunbar, Jill. "NAS Parallel Benchmarks". NASA. 2012. http://web.ipac.caltech.edu/staff/fmasci/home/miscscience/MIC_benchmarking_2013.pdf

[7] Jin, H and Frumkin, M and Yan, J. "The OpenMP Implementation of NAS Parallel Benchmarks and Its Performance". 1999. https://www.nas.nasa.gov/assets/pdf/techreports/1999/nas-99- 011.pdf

[8] Roth, Frances. "System Administration for the Intel Xeon Phi Coprocessor". Intel. 2013. http://software.intel.com/sites/default/files/article/373934/system-administration-for-the-intel-xeon-phi-coprocessor.pdf

[9] Green, Ronald. "OpenMP Thread Affinity Control". Intel Developer Zone. 2012. http://software.intel.com/sites/default/files/article/373934/system-administration-for-the-intel-xeon-phi-coprocessor.pdf

[10] Vaibhav, Sundriyal and Sosonkina, Masha. "Initial Investigation of a Scheme to Use Instantaneous CPU Power Consumption for Energy Savings Format". Proceedings of the 1st International Workshop on Energy Efficient Supercomputing. 2013.

Business, Industry, Infrastructure Security, & Military

VMASC Track Chair: Dr. Saikou Diallo, Dr. Barry Ezell
MSVE Track Chair: Dr. Jim Leathrum, Dr. Bharat Madan, Dr.ManWo Ng

Deep Model for Improved Operator Function State Assessment
Author(s): Feng Li, Jonathan Wen, Jiang Li, Guangfan Zhang, Roger Xu, and Tom Schnell

Decision Points – Laying a Foundation for Vehicle and Pedestrian Interactions
Author(s): Terra Elzie

Discrete Event Simulation Implementation of a Production Planning and Scheduling Tool
Author(s): Jesse Cladwell, Christopher Heard, Ashton Allen, Ioannis Sakiotis, and Daniel Drake

Afghanistan and US bargaining over Bilateral Security Agreement
Author(s): Khatera Alizada

Water Security in the Kabul-Kunar River Bason
Author(s): Amanda Norton

Deep Model for Improved Operator Function State Assessment

[1]Feng Li, [2]Jonathan Wen, [1]Jiang Li, [3]Guangfan Zhang, [3]Roger Xu and [4]Tom Schnell
[1]Dept. of Electrical and Computer Engineering, Old Dominion University, Norfolk VA
[2]The IB Program at Princess Anne High School, Virginia Beach, VA 23456
[3]Intelligent Automation Inc., Rockville, MD
[4]Department of Industrial Engineering, University of Iowa, Iowa City, IA

***Abstract* – A deep learning framework is presented for engagement assessment using EEG signals. Deep learning is a recently developed machine learning technique and has been applied to many applications. In this paper, we proposed a deep learning strategy for operator function state (OFS) assessment. Fifteen pilots participated in a flight simulation from Seattle to Chicago. During the four-hour simulation, EEG signals were recorded for each pilot. We labeled 20-minute data as engaged and disengaged to fine-tune the deep network and utilized the remaining vast amount of unlabeled data to initialize the network. The trained deep network was then used to assess if a pilot was engaged during the four-hour simulation.**

***Index Terms* -- Deep learning, Engagement assessment, Physiological signals.**

I. INTRODUCTION

The Operator Function State (OFS) observes the performance of an individual when placed in high stress scenarios that test both physiological and psychological strength. While a high load of tasks can be incredibly strenuous and prove demanding, a small load of tasks can prove just as dangerous as this leaves the person not focused and could possibly yield bad decisions. As a result, it is important to assess OFS to make sure a person stays focus. In particular, pilots must face various scenarios which range from unexpected events to emergency scenarios; moreover, a poor decision can result in fatal consequences. In order to investigate the situation, most OFS assessments observe the psycho-physiological patterns that are present and associate those patterns with cognitive states including fatigue, engagement, and others [1].

This paper uses a deep learning framework [2] to extract effective patterns/features from EEG signal for OFS assessment. We invited 15 pilots to participate a four-hour flight simulation from Seattle to Chicago and recorded their EEG signals during the entire simulation. The deep learning framework was then utilized to extract effective features from the EEG signals to identify if the pilots were engaged during the simulation.

II. Proposed Methods

A. *Data Collection*

To study the engagement of pilots, a Boeing 737 simulator as shown in Fig. 1 was used to simulate an actual four-hour flight from Seattle to Chicago. Fifteen pilots were invited in the experiment conducted at the University of Iowa. The data collected include technical data and EEG. Twenty-minute data for each pilot was labeled as different OFS states (engaged and disengaged) and the remaining data was left unlabeled. Using the data collected, we investigated OFS assessment for the pilots through three modules based on the deep learning framework. The modules include EEG signal processing, pre-training of the deep network with unlabeled data, and fine-tune the network with the labeled data. These shall be further explained in the following subsections.

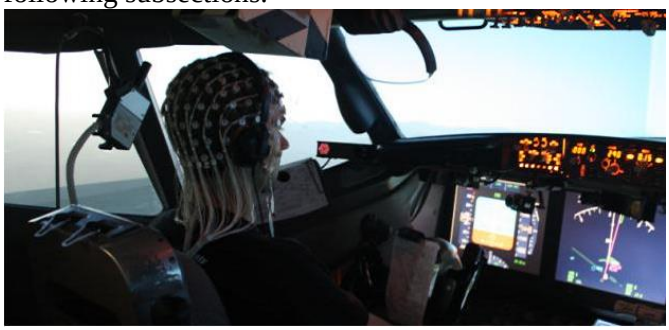


Figure 1. Flight Simulator.

B. EEG Signal Processing

By following the rules for bi-polar site configuration for engagement assessment, EEG data was collected through 8 pairs of EEG sensor nodes for a total of 32 sensors. However, these signals are known to be contaminated and filled with physiological and non-physiological signals that deter from the actual data. In this paper, we first removed the unusual patterns such as spikes and saturated areas. The signals were then divided into three second segments and we extracted 39 power spectral features for each segment which brings a

grand total of 312 features since there were 8 channels used. Based on the 312 features, we trained a deep network (described in the next subsection) to assess if the pilots were engaged. The three second time window then was moved forward with a step size of one second to make another assessment.

C. *Pre-training*

We utilized a deep model of 4 hidden layers with the number of hidden units as 200, 100, 50, and 20 as shown in Fig. 2 for OFS assessment. The unlabeled data was used to pre-train the hidden layers once at a time in a layer by layer manner. This procedure was based on the restricted Boltzmann machine (RBM) and is unsupervised [2], yielding four layers of pre-trained hidden layers.

D. *Fine-tuning of the network with lablled data*

After pretraining the deep network, we utilized partially the 20 minutes labeled data to fine-tune the network through the use of a back propagation algorithm [3]. The final trained network was then applied to the remaining labeled data to evaluate the effectiveness of the trained deep model.

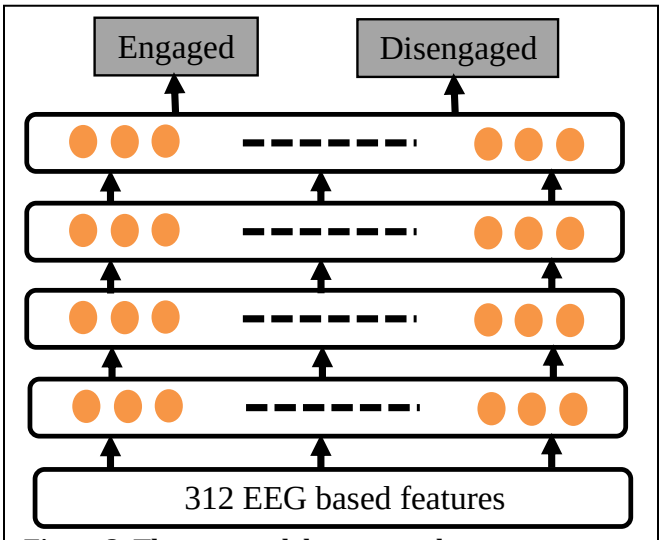


Figure 2. The proposed deep network.

III. Experiments and Results

For the 15 pilots, we created a separate deep model to assess if the pilot was engaged during the simulation. For each pilot, the vast amount of unlabeled data was used to pre-train the deep model and various amounts of the labeled data was applied to fine-tune the model. Once the deep model was finalized, we applied it to the remaining labeled data from the pilot to obtain an accuracy of the model for engagement assessment. It has been shown that the pre-training step in deep learning is critical for its success [2]. In this paper, we compared the deep model with and without the pre-training step for engagement assessment. .

Table 1 shows the averaged experimental results for the 15 pilots with and without the pre-training step. We utilized different amount of labeled data for fine-tuning the deep model, for example, 3% means that only 3% of the labeled data was used for fine-tuning. If a deep model was obtained without pre-training, its initial weights were randomly generated and thus the deep model is similar to a traditional multilayer perceptron.

Table1: Engagement assessment accuracy (%) averaged across 15 pilots. X% stands for using X% of the labeled data to fine-tune the deep model.

Scenario	%3	%5	%10	%20	%50
Pre-trained	87.86	89.22	89.42	90.10	92.40
No pre-train	83.92	86.93	87.90	89.11	91.66

It is observed from Table 1 that the proposed deep model can effectively predict if a pilot was engaged during the simulation. If the deep model was pre-trained with the RBM mechanism, the model will be much more effective for OFS assessment if labeled data is limited. Using only 3% of the labeled data for fine-tuning, the pre-training step can boost the performance from 83.92% to 87.86%. This is particularly important because labeling the data to different OFS states is not only technically difficult but also time consuming.

IV. Conclusion

We proposed a deep learning strategy for OFS assessment in this paper. In particular, we invited 15 pilots to participate in a flight simulation to assess if they were engaged in the simulation. The assessment was based on power spectral features extracted from EEG signals recorded during the simulation. We also showed that the pre-training step in deep model is critical for the engagement assessment especially when labeled data is limited.

Reference

[1]. R. Hockey, "Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks," *NATO ASI SERIES Vol 355.*

[2]. G.E. Hinton, S. Osindero and Y. Teh, “A fast learning algorithm for deep belief nets”, Neural Computation, 18, pp 1527-1554, 2006.

[3]. D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning representations by back-propagating errors". Nature, 323 (6088): 533–536, 1986.

Decision Points – Laying a Foundation for Vehicle and Pedestrian Interactions

T. Elzie, Old Dominion University, VMASC Scholar

***Abstract* – Simultaneous evacuation of vehicles and pedestrians is a realistic scenario for many types of venues such as egress after sporting events, concerts, or mega-church services. Disaster managers, emergency planners, and local government decision-makers benefit from the use of simulation models. Therefore, developing an accurate and realistic model that incorporates vehicle and pedestrian interaction is essential. This paper builds on phase one of an existing pedestrian model that analyzes group dynamics during an evacuation, extending it to explore vehicle and pedestrian interactions in the same network using an agent-based modeling approach. This conceptual model details the decision-making perspectives of both vehicle drivers and pedestrians. First, the preferred destinations are briefly discussed; then the various decision points for pedestrians and vehicle drivers are detailed using UML activity diagrams. This model will produce improved results for simultaneous evacuations of vehicles and pedestrians by incorporating heterogeneous behaviors and characteristics for each agent while capturing five different types of interactions: 1) normal, 2) controlled, 3) random, 4) chaotic concordant, and 5) chaotic conflicting [1].**

***Index Terms* – Agent-based, Evacuation, Interaction, Pedestrian, Simulation, UML, Vehicle**

I. INTRODUCTION

Utilizing simulation models for simultaneous evacuation of vehicles and pedestrians is an essential analytical tool for disaster managers, emergency planners, and local government decision-makers. Developing an accurate and realistic model that incorporates vehicle and pedestrian interaction, however, is an ongoing challenge, and continues to be an area of interest to researchers.

Currently, the author is working as part of a transportation project team that is creating a pedestrian behavior agent-based model (ABM) to simulate evacuation of venues that have no vehicle presence. The main crux of this project is to incorporate inter- and intra-group dynamics to extend mainstream pedestrian models beyond the individual level of analysis. The autonomous, heterogeneous nature of agents in ABMs allows for modeling of group dynamics that more realistically capture the way crowds composed of families and other social groups egress during an emergency. The next development phase of this project will introduce vehicles to address pedestrian-vehicle interactions within the same network, thus introducing an element that is scarcely present in existing evacuation simulations.

Building toward the second development phase of this project, this paper describes a conceptual model that captures the decision-making process during interactions between pedestrians and vehicle drivers while egressing during a controlled or emergency evacuation. Where the first phase of the project strictly centered on group dynamics and decision-making behavior among interacting pedestrians, this conceptual model incorporates vehicle and driver behavior to understand how this affects overall evacuation outcomes. For simplicity, this paper considers only the behaviors of autonomous pedestrian agents rather than those considering group behaviors. In subsequent phases, once

individual pedestrian and vehicle interaction are thoroughly modeled, group dynamics will be reintroduced as a critical part of a full, realistic model.

ABM was selected to provide a more representative approximation of real-world crowd egress and vehicle evacuation. Other major approaches to simulating pedestrian and vehicle behavior is cellular automata (CA) models and social force dynamic models.

Cellular automata (CA) requires discrete space. Agents vary in level of sophistication from those that follow simple rules to those that dynamically learn from their environment. CA is characterized as an artificial life approach to simulation modeling and is named after the principle of automata (entities) occupying cells according to localized neighborhood rules of occupancy [2]. The CA local rules prescribe the behavior of each automaton creating an approximation of actual individual behavior.

Agents in social force dynamic models incorporate path-finding algorithms and perception control theory to try to control their relationships with others. Helbing and Molnar [3] pioneered this methodology by likening social forces acting on an agent to physical forces. These forces are then combined with other forces to produce a resultant force (direction and speed) for each agent.

Bonabeau [4] summarized the benefits of ABMS over other modeling techniques as follows:

- ABMS captures emergent phenomena.
- ABMS provides a natural description of a system.
- ABMS is flexible.

Specifically, ABMS is superior in modeling the following situations:

- The interactions between agents are complex, nonlinear, discontinuous, or discrete.
- Space is crucial, and agents' positions are not fixed.
- Population is heterogeneous, and each individual possesses different characteristics.
- The topology of the interactions is heterogeneous and complex.
- Agents exhibit complex behavior, especially involving learning, interactions, and adaptation.

A critical feature of agents in ABM include not only learning and heterogeneity, but also the ability to learn from interactions with other agents.

Section 2 reviews the existing literature for research on vehicle and pedestrian interactions. Section 3 discusses a UML-based conceptual model to work through the most important components of a vehicle-pedestrian ABM. The final section concludes with thoughts on how this pedestrian ABM will ultimately be of practical use to disaster managers, emergency planners, and local government decision-makers.

II. LITERATURE REVIEW

This section presents an overview of how vehicles and pedestrians typically interact in areas of conflict in order to properly develop this multimodal environment. Gentile et al. [1] define five different types of vehicle-pedestrian interactions progressing from the best case to worse case: 1) normal, 2) controlled, 3) random, 4) chaotic concordant and 5) chaotic conflicting. A description of each is given below:

- *Normal Interactions* occur in ordinary scenarios, where pedestrians use only sidewalks and cross the street at intersections. The vehicle-pedestrian interactions are neglected and only vehicle turn delays are considered. However, pedestrian discordant or contra-flow interactions must be considered on the sidewalks.

- *Controlled Interactions* occur during special events, where some road lanes are assigned to pedestrians and do not spread to lanes reserved for vehicles. It is assumed that no longitudinal interactions will occur when sidewalks are highly crowded and some faster moving pedestrians occupy parts of the road lane nearest the sidewalk. However, due to high pedestrian volumes, transversal interactions cannot be avoided when pedestrians cross the street causing the vehicle flow to be delayed as they yield to the pedestrians.

- *Random Interactions* occur during special events and/or evacuation scenarios. Although pedestrians should stay on sidewalks (or assigned lanes), they may randomly occupy part of the

vehicle lane nearest the sidewalk, causing the vehicle-pedestrian longitudinal interaction.

- *Chaotic Concordant Interactions* occur during an evacuation scenario where it is impossible to separate pedestrian and vehicle flows mixed completely together. However, on each road, pedestrian and vehicles flow in the same direction. The assumption here is that during the evacuation, both vehicles and pedestrians know the direction/ location of evacuation/ collecting point, so everyone is trying to reach the same point. The sidewalk capacity is assigned to pedestrians while the entire road capacity is assigned to both vehicles and pedestrians, thus vehicles are forced to travel at pedestrian speed.

- *Chaotic Conflicting Interactions* are the worst-case evacuation scenario where it is impossible to separate pedestrian and vehicle flows. Both are completely mixed together and pedestrians do not respect vehicle directions. Resulting in vehicles 'stuck' on the road while pedestrians flow around them.

These five areas categorize vehicle-pedestrian interactions in a way that facilitates development of an accurate and complete simulation model.

The existing literature on pedestrian simulation models provides relatively little evidence of pedestrian-vehicle interactions. Those that capture this feature model very narrow aspects of the interaction. Some existing models acknowledge the importance of considering pedestrian behaviors in vehicle networks, but the modeling paradigm is too abstract to capture individual level behaviors or characteristics in order to understand the effects of these factors on the evacuation outcomes. For example, Rossetti and Ni [5], present a microscopic simulation model of large-scale evacuations of parking lots in a commercial shopping district. Through this model, they explore the effects that exiting vehicles have on the surrounding traffic flows in the network. Although they acknowledged that pedestrian interferences would play an important role in affecting the evacuation time, particularly within the parking lot, specific aspects of conflicting vehicle-pedestrian entities was not taken into account. The authors acknowledge this lack of detail causes their simulation to be deficient. They concede that incorporation of the vehicle-pedestrian interaction in the parking lots would improve the overall evacuation and its effect on the background traffic.

Meschini and Gentile [6] simulate vehicle-pedestrian interactions during mass events using a macroscopic dynamic traffic assignment (DTA) model that calculates dynamic user equilibrium. They extend this approach to represent pedestrian flows and vehicle-pedestrian interactions. However, the road network, pedestrian network and sidewalk network were represented separately in the model, with relationships drawn between them.

Sun and Benekohal used a microscopic simulation approach that focused on vehicle-pedestrian interaction at uncontrolled mid-block crosswalks, a model they call the Pedestrian Motorist Interaction Simulator (PMIS) [7]. The simulator is a hybrid model that consists of the combination of the car-following logic to move the vehicles and three additional models to represent the decision-making process of the motorists and the pedestrians. The three additional models are the Motorist Yield (MOY), Pedestrian Gap Acceptance (PGA) and the System Dynamic Interaction (SDI) models [7]. Sun and Benekohal used a binary logit method to build the model that considers multiple variables for pedestrians and vehicles when faced with a conflicting scenario within the network.

Although these models and techniques are useful in certain contexts, the goal of our simulation model is full integration of vehicle-pedestrian interactions using an agent-based model approach. By incorporating certain behaviors and characteristics for each agent, while capturing the five different types of interactions as mentioned above, this model will produce improved results for simultaneous evacuations of vehicles and pedestrians.

III. CONCEPTUAL MODEL

Individual pedestrian and vehicle behaviors and characteristics established in the conceptual model allow decision-making to take place as agents encounter conflicts. Conceptualizing this through UML diagrams defines destinations and decision points for vehicles and

pedestrians as interactions occur throughout the model. The UML diagrams described in this paper focus only on the decision points for vehicles and pedestrians. To be all-encompassing, future versions of the model will capture vehicle-pedestrian interactions as well as pedestrian-pedestrian interactions.

Pedestrian and Vehicle Destinations

From the pedestrian perspective, the destination of a person departing a venue is two-fold. First, if inside a building, the destination is the preferred exit to leave the building. Clearly, in this scenario, all interactions are pedestrians with other pedestrians. Second, once outside, the pedestrian must now reach a final destination. This destination can vary, for example home, a bus stop or a light rail station, or reaching their vehicle. In this scenario, pedestrians interact with other pedestrians as well as with vehicles. Once an individual reaches his or her vehicle, the destination is updated; this destination can be tracked all the way home, for example. However, for purposes of egressing a venue, the destination can be thought of as reaching a point on one's driving path that is beyond the mass exit of pedestrians and pedestrian conflicts, thus permitting the vehicle to reach the allowable driving speed outside of a predetermined radius from the venue.

Pedestrian Decision Points

The ultimate goal of a pedestrian is to reach his or her final destination. However, various decision points, both inside and outside the venue, exist between the current location and final destination. Figure 1 and Figure *2* visually represent decision-making processes required to navigate through the crowd to a preferred destinations.

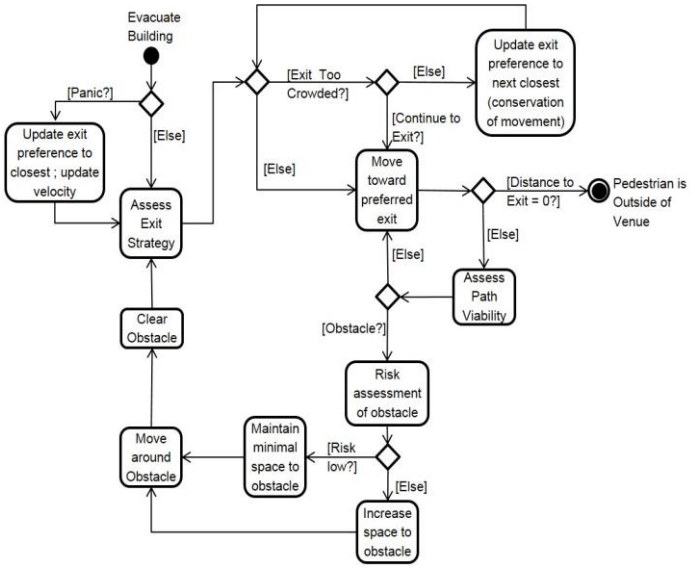


Figure 1: Pedestrian Decision Points Inside of Venue

Since pedestrians are individual agents that are autonomous and possess heterogeneous traits, each pedestrian makes different decisions than other pedestrians based on their varying characteristics. For example, an agent may be characterized as being a 'leader' or exhibiting individualistic behavior [8] suggesting that when navigating through a crowd, this agent would determine his own path and not consider what other agents are doing around him. On the other hand, an agent may be characterized as a 'herd follower' [8] [9], where the path is determined by influences of a crowd flowing in the same general direction.

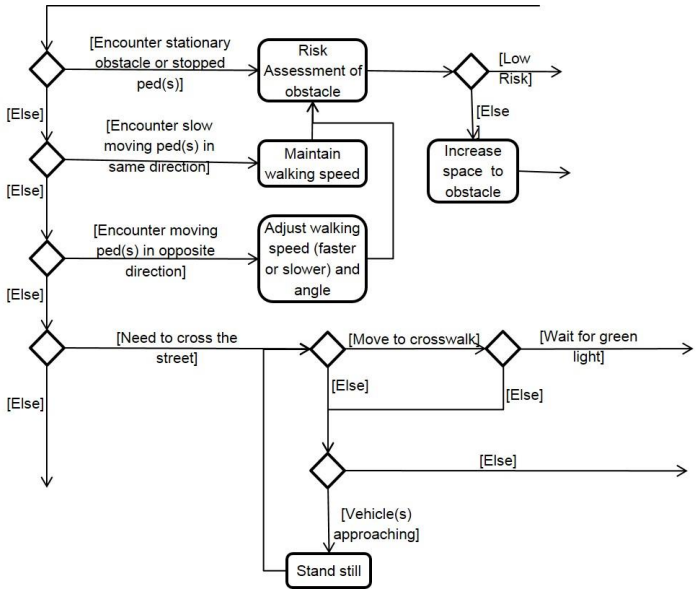


Figure 2: Pedestrian Decision Points Outside of Venue (Partial)

Whether a pedestrian is inside a building or outside, similar decision points are required. Inside a building for instance, if a preferred exit is too crowded, the agent would decide to either wait to exit or change course to the next closest exit. For outside the venue, if the path to the preferred final destination (vehicle, bus stop, etc.) is too crowded, then avert the crowd and maneuver to another viable path with less crowd density. Stationary and moving obstacles present another decision point for pedestrians. An obstacle could be a kiosk or parked vehicle, a pedestrian or group of people standing still, a slower moving pedestrian walking in the same direction, or a pedestrian walking directly toward another. The agent conducts a risk assessment of the obstacle in order to determine the best course of action to maneuver pass the obstacle. Once assessed, an agent may increase or decrease walking velocity or change the walking angle to pass the obstacle. Another maneuver may be to increase the distance from the obstacle if the risk of danger is too high.

Crossing a street is another major decision point for a pedestrian in order to continue navigating to the final destination as shown in Figure 2. Certain characteristics of an agent will contribute greatly to how and when an agent crosses the street. An "obedience level," capturing an agent's likelihood to obey pedestrian laws, factors into this decision. Agents' risk aversion levels determine their likelihood to abide by crosswalks and pedestrian signaling. Any type of jaywalking (crossing midblock or crossing a crosswalk during a red signal) will present a potential conflict with oncoming moving vehicles, requiring the pedestrian to make decisions about risk presented by the vehicle's speed and his or her own mobility level.

Vehicle Decision Points

Vehicle decision points occur at three instances: 1) car following (when to accelerate, decelerate, or stop), 2) encountering pedestrians mostly in a jaywalking situation, and 3) at intersections (turning right, left or continue straight) where pedestrian interaction is possible. Figure 3, Figure *4*, and Figure *5* show the details of these three decision points, respectively, as a vehicle is traveling toward the final destination. These three figures combine into one diagram. This model only incorporates one lane for vehicles; therefore, no lane changing is represented in this preliminary concept.

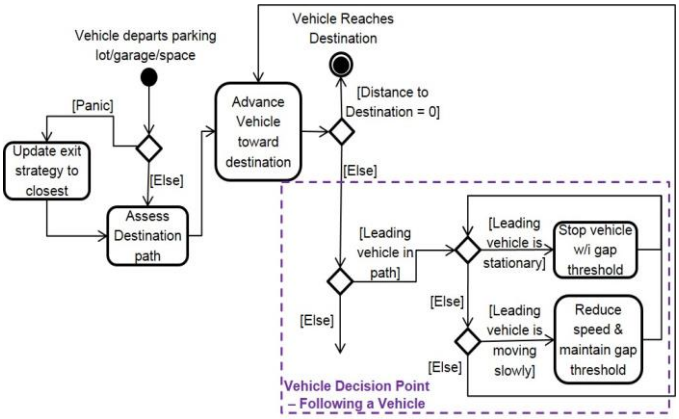


Figure 3: Vehicle Following Vehicle Decision Point

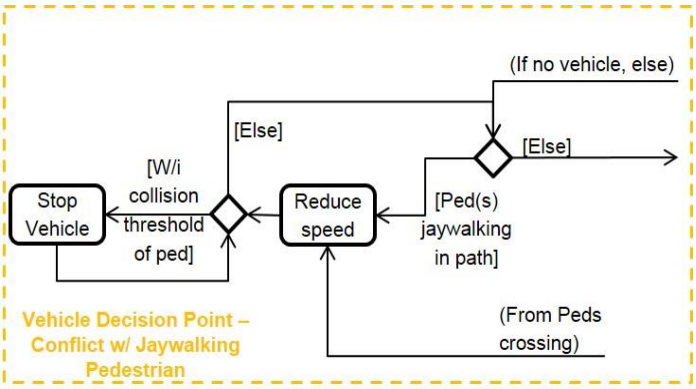


Figure 4: Vehicle Conflicting w/ Pedestrian Decision Point

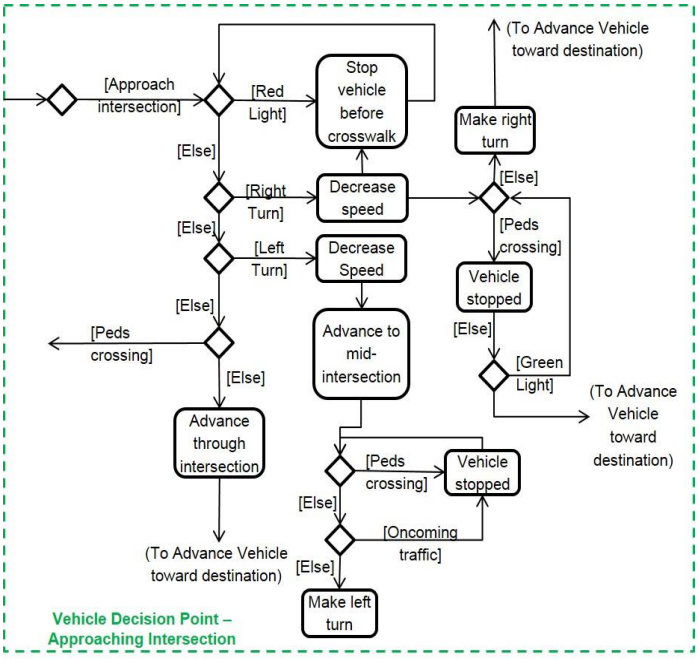


Figure 5: Vehicle Approaching Intersection Decision Points

The potential interactions with pedestrians arise mainly when pedestrians decide to jaywalk either at intersections when crossing on red or at midblock. Depending on the characteristics of the vehicle driver, such as level of aggression while driving, the decision-making during vehicle-pedestrian conflicts will vary.

IV. CONCLUSIONS

This paper presents a conceptual model that extends a current study of pedestrian-pedestrian interactions by incorporating vehicle-pedestrian interactions during evacuation scenarios. Using an agent-based modeling approach allows for observing driver and pedestrian behavior as these autonomous and heterogeneous agents make decisions throughout the model based on their individual characteristics. As described in this paper, each agent, no matter the mode of transportation, has an origin and a destination. This conceptual model highlights each specific decision point along agents' evacuation paths using UML activity diagrams. As the research team develops the simulation model, implementation of the five different types of vehicle-pedestrian interactions (normal, controlled, random, chaotic concordant, and chaotic conflicting) into the model will equip the simulation with a high level of realism that represents more accurate simulation results from controlled to emergency evacuations. In addition, using agent based modeling, allows for capturing emergent behavior that may arise in the overall system due to the various interactions among pedestrian and vehicle agents, exposing unexpected evacuation results. Once complete, the integration of the pedestrian group dynamics with vehicle-pedestrian interactions will facilitate more informed planning by disaster managers, emergency planners, and local government decision-makers.

V. REFERENCES

[1] G. Gentile, L. Meschini and N. Papola, "Spillback congestion in Dynamic Traffic Assignment: A Macroscopic Flow Model with Time-Varying Bottlenecks," in *Transportation Research Board*, 2007.

[2] S. Levi, "Artificial Life," New York, Vintage Books, 1992.

[3] D. Helbing and P. Molnar, "Social Force Model for Pedestrian Dynamics," *Physical Review*, 1995.

[4] E. Bonabeau, "Agent-Based Modeling: Methods and Techniques for Simulating Human Systems," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7280-7287, 2002.

[5] M. D. Rossetti and Q. Ni, "Simulating Large-Scale Evacuation Scenarios in Commercial Shopping DISTRICTS – Methodology and Case Study," in *Proceedings of the Winter Simulation Conference*, 2010.

[6] L. Meschini and G. Gentile, "Simulating Car-Pedestrian Interactions During Mass Events with DTA Models: The Case of Vancouver Winter Olympic Games," in *European Transport Conference*, Netherlands, 2009.

[7] D. Sun and R. F. Benekohal, "Modeling and Simulation of Pedestrian-Motorist Interaction at Uncontrolled Mid-block Crosswalks," in *Institution of Transportation Engineers (ITE) Technical Conference and Exhibit*, Fort Lauderdale, 2003.

[8] D. Helbing, I. Farkas and T. Vicsek, "Simulating Dynamic Features of Escape Panic," *Nature*, pp. 487-490, 28 September 2000.

[9] A. Schadschneider, A. Kirchner and K. Nishinari, "CA Approach to Collective Phenomena in Pedestrian Dynamics," in *Cellular Automata*, Springer Berlin Heidelberg, 2002, pp. 239-248.

Discrete Event Simulation Implementation of a Production Planning and Scheduling Tool

Jesse Caldwell, Christopher Heard, Ashton Allen, Ioannis Sakiotis, and Daniel Drake
Department of Modeling, Simulation, and Visualization Engineering
Old Dominion University
[jcald013, chear008, aalle041, isaki001, ddrak006]@odu.edu

Keywords: Discrete Event Simulation, Process Flow, Job Shop, Scheduling-

Abstract: This paper describes the implementation of a discrete event simulation production planning and scheduling (PPS) tool for use in a job shop manufacturing facility. The PPS tool's architecture is comprised of four main sub-systems, the system input, PPS simulation, system output, and controller. Each of those sub-systems is independent and can only communicate through a set of pre-defined interfaces. This paper provides a high-level description of the PPS architecture with an emphasis on the design and implementation of the PPS simulation component.

INTRODUCTION

Background Information

The Modeling, Simulation, and Visualization Engineering (MSVE) Senior Capstone Design Team (SCDT) from Old Dominion University (ODU) is working with Newport News Industrial (NNI) and Newport News Shipbuilding (NNS) to create a prototype software tool that will assist NNI with their production planning and scheduling process (PPS). This prototype is an adaptive PPS simulation based tool that is being developed for NNI. The SCDT is receiving support from NNS's Modeling and Simulation department throughout the lifetime of this project.

The SCDT is participating in a two semester Capstone Design course which is mandatory for students majoring in MSVE at ODU. The purpose of this course is to prepare students for the professional workforce. In this course the students work with an outside customer to complete a project. The students go through the engineering design process to define the problem, design a solution, and implement the design. The SCDT has observed NNI's planning and scheduling process and has designed a solution that will result in a prototype simulation tool. This software is being designed so that NNI and NNS can continue development after the delivery of the prototype PPS simulation tool.

NNS is a division of Huntington Ingalls Industries (HII) and NNI is a subsidiary of NNS. The Modeling and Simulation department of NNS has a unique role in this project; NNS has asked the SCDT to use the Common Simulation Framework (CSF). NNS develops and maintains this Java-based framework within the department. Throughout the project, NNS is supporting the SCDT by acting as a consultant for the component development that requires CSF.

The primary customer for this project is NNI's Oyster Point Industrial Park (NNI-OP) facility. NNI-OP is a unique manufacturing plant that undertakes a variety of complex jobs. NNI-OP's core capabilities consist of machining, welding, rigging and mechanical assembly. This wide range of capabilities results in a variation of job requirements that make scheduling difficult. This paper provides a high-level overview of the complete system architecture with an emphasis on the design and implementation of the PPS simulation component [1].

Paper Organization

The paper is organized as follows: Problem Definition, System Overview, Simulation Design, Simulation Sub-System Design, Future Direction, and the Conclusion. The Problem Definition describes NNI-OP's current scheduling process and explains how the PPS tool will support NNI-OP's scheduling decisions. The System Overview describes the system architecture at a high level and presents the approach being taken to design the system. The Simulation Sub-System provides a detailed description of the design of the simulation sub-system. The Future Directions section explains how the simulation sub-system will be implemented using the CSF. The Conclusion states the current stage of the development and identifies features that could be added after prototype delivery.

PROBLEM DEFINITION

Current Planning Process

When new work is proposed, the planners must evaluate the requirements for a new job, and determine how that job will impact the current schedule of the facility. The planners determine the requirements for a new job by using past job experience. The planner then considers the effects that the new job will have on the current state of the facility and the schedule of work to be completed [2]. NNI-OP generates a document that is intended to assist with the scheduling process. This document, known as a Job Tracking Sheet, represents each job as a sequence of tasks. Each task has an expected processing time and a list of resources that are required to complete the task.

Problem Description

Currently, each job is created and documented independently. This means that there is no way to easily visualize the jobs that are scheduled in the facility or the effects of inserting a new job into the existing schedule. The information on the Job Tracking Sheet does not show the status of the facility, the possibility of rework, or the time a task spends waiting for a resource.

The proposed tool will give the planners the ability to simulate a set of jobs in the facility. This simulation approach accounts for resource contention and the possibility of rework. The user will be able to make confident decisions based on the output provided by the tool through job schedules, machine schedules and performance statistics.

PPS SYSTEM OVERVIEW

This simulation consists of four major sub-systems: the System Input, the PPS Simulation, the System Output, and the Controller. Between these sub-systems are pre-defined interfaces that allow them to communicate. The high-level system architecture is shown in Figure 1.

System Input

The system input is represented by the blue section in Figure 1. The system input consists of three sub-systems: Job File Developer, Scenario Developer, and the Resource Database. The Job File Developer allows the user to input information that describes the tasks that define a job. The job information is then saved and stored in the Job Repository. A job is represented as a single job file. When the job files are developed, the resources required by a task are checked against the resources available in the resource database. The resource database contains information about all of the resources available within the facility: workers, machines, and workstations. The scenarios are a collection of jobs that are assigned start dates and priority. Once a scenario is developed it is stored in the scenario repository where it can be accessed by the PPS Simulation partition to be run through the simulation.

PPS Simulation

The PPS simulation consists of three components, the Initialization Module, Task Flow Processor (TFP), and Resource Processor (RP). The initialization module reads in the scenario file and the job files associated with the scenario file. Then the model is initialized and set up by reading in the resource file. The task flow processor manages each job, runs the simulation, and generates output that is stored in the output repository. The resource processor manages the facility resources in the simulation by allocating resources in response to the requests from the task flow processor.

PPS Output

The system output handles all of the data analysis and visualization for the simulation. It consists of six sub-systems: the Single Run Analyzer, Multiple Run Analyzer, Schedule Repository, Performance Repository, Schedule Display, and Performance Display. The Single Run Analyzer reads data from the Output Repository for a specific run and organizes the data into a format that the Schedule Display can read and display. The Schedule Display retrieves data from the Schedule Repository and generates a Gantt chart display of job schedules. The Multiple Run Analyzer collects the data from the Output Repository and batches it together. This data is then used to compute statistical estimators for quantities such as expected completion date. Once these statistical estimators have been computed, the Multiple Run Analyzer sends the data to the Performance Repository. The Performance Display reads the data from the Performance Repository and presents the results in a convenient visual display.

Controller

The Controller is a sub-system which serves as the communication method between the various sub-systems and the user. The Controller contacts the System Input and Output sub-systems whenever they needed to be launched. The Controller interacts with the PPS simulation by passing various simulation parameters such as number of replications. The Controller provides to the PPS simulation a reference

to the scenario that is to be run. Then the Controller executes the PPS simulation.

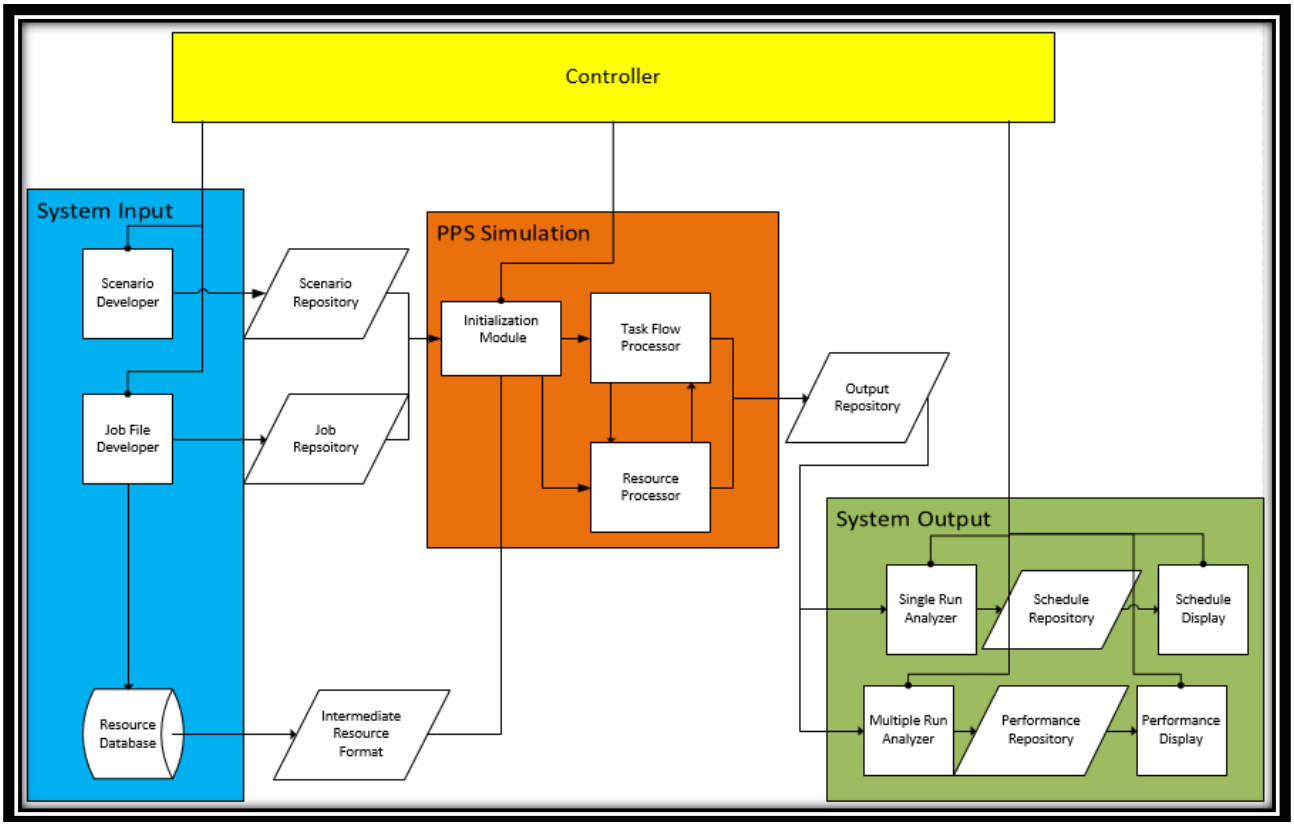


Figure 1. High level design

Component Interfaces

The interfaces serve as communication platforms between sub-systems of the PPS tool. These interfaces allow each of the sub-systems to be created independently as long as the developers remain faithful to the specified interface format. This allows for any new developer to alter or substitute any sub-system.

SIMULATION DESIGN

The simulation design was develop by decomposing NNI-OP's scheduling processes into the three sub-components of the PPS Simulation: Initialization Module, Task Flow Processor, and Resource Processor. This can be seen in Figure 1. The next sections expand on Figure 2 and show the logic and design of how these sub-components interact to form the functionality of the PPS simulation.

Initialization Module

The Initialization Module consists of two main components and two sub-components: the Scenario File Reader, the Model Initialization, as well as, the Job file Reader and the Resource File reader, respectively. The Scenario File Reader reads the scenario file that is provided by the Controller and models a job task flow for each type of job. If any job has a task flow that matches another job's task flow, only one job type is created in order to preserve memory. Once the job task flows have been created, the Model Initialization component calls the Resource File Reader which adds all specified resources to the RP. Once all of the resources are finalized, the simulation begins.

Task Flow Processor

The TFP determines the current task for a specific job and manages job arrivals and departures from the simulation. The TFP also collects data regarding the statistical characteristics of each task and forwards it to the Output Repository. Each task has a variety of outcomes that need to be accounted for when navigating to a new task. For each transition to a new task, three unique cases are accounted for within the TFP. The first case occurs when there is only a

single task that linearly follows the current task. The new active task replaces the current task as the new active task and is sent to the RP to be completed. The second case occurs when there are two or more tasks that follow the completed task and the probabilities of each task occurring must sum to one. Tasks that exhibit this behavior are referred to as dependent conditional tasks and often represent an inspection task which can yield two outcomes, pass or fail. The third case occurs when

there are two or more tasks proceed the completed task and the probabilities need not sum to one. Tasks that exhibit this behavior are referred to as independent conditional tasks. These tasks appear when a larger item, such as a valve, is broken down into its components that can be worked on simultaneously in the facility.

Resource Processor

The RP is where the resources reside and is responsible for resolving resource contention between tasks. The RP also collects data on how long a task is at a given resource, the value added and non-value added time a task spends at a resource and resource utilization. It then forwards this information to the Output Repository.

The RP consists of two major types of sub-components: the Resource Router and the individual resource models. Each resource model comprised of two sub-components: a queue and a workstation. When the resource router receives a task from the TFP, which can be processed by different resources, it finds the resource with the lowest expected delay E_D as shown below;

$$E_D = d + \sum d_k$$

where d_k is the expected processing time of the k_{th} element with a priority greater than or equal to the task's priority in the resource's queue and d is the expected processing time of the current task. Once E_D has been computed for each resource, the Resource

Router sends the task to the resource with the smallest E_D.

SIMULATION SUB-SYSTEM DESIGN

There are two components in the PPS Simulation sub-system, the TFP and the RP. Each of these components is comprised of various sub-components.

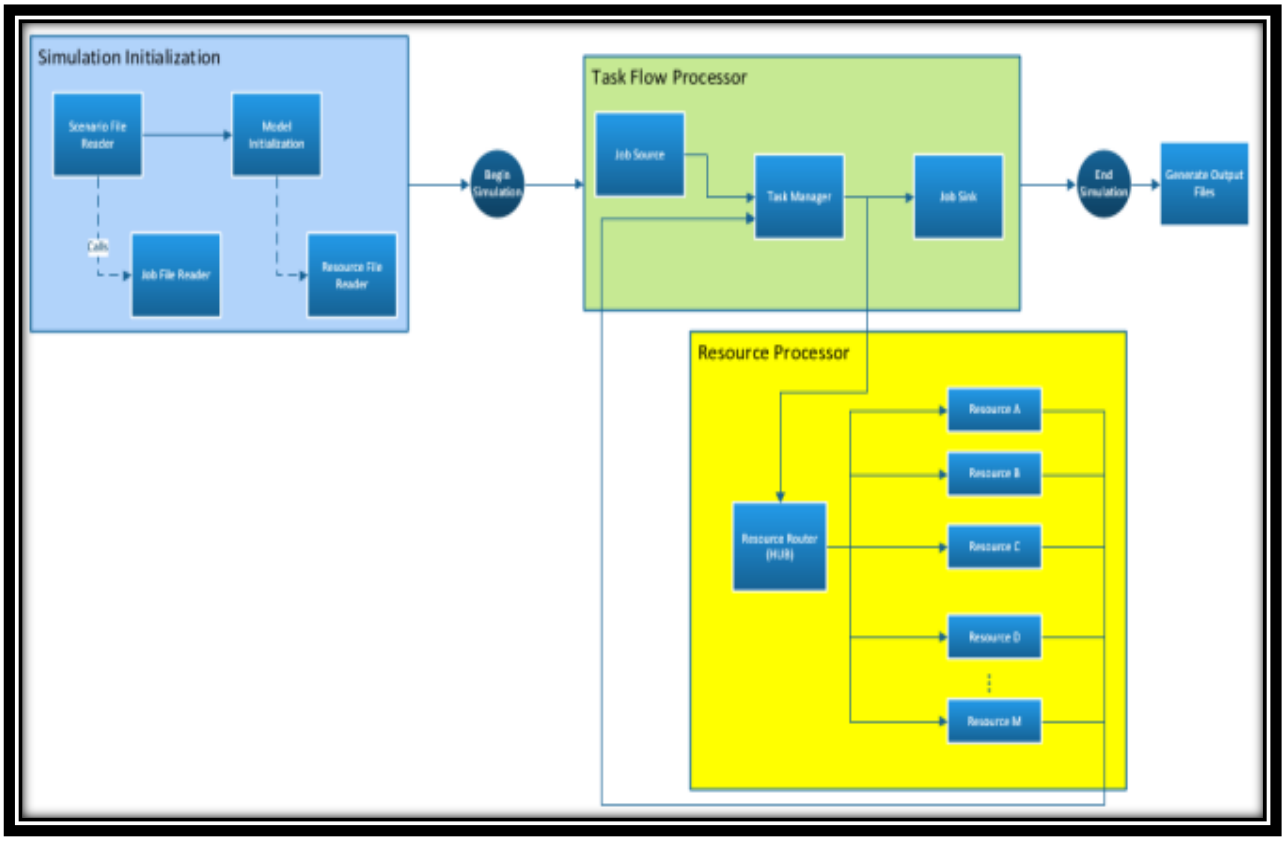


Figure 2. PPS Simulation Sub-system Breakdown

Task Flow Processor Design

The TFP is made up of three components: the job source, the task manager, and the job sink. The job source and job sink components is represented as a source and a sink, as their name implies. A source and sink creates and removes an entity within the simulation respectively. The task manager component is represented by a series of processors and routers.

The first router sends an entity to the proper processor, based on the type of task the entity has just completed. If the completed task was a process, the entity is sent to a processor that assigns the next task to the entity. If the completed task was an inspection, the entity is sent to the other router, which sorts the entity by type of inspection, dependent or independent. If the task was a dependent task, it is sent to a processor that determines the result of the inspection (pass or fail) and assigns the next task to the entity, based on the result of the inspection. If the task was an independent task, it is sent to a processor which determines how many instances of the entity will be created. This is based on the number of simultaneous tasks the job must complete to exit the facility.

Resource Processor Design

The RP is composed of two components: the resource router and the various resource models. The resource models are made up of a priority, FIFO queue and a workstation, to represent to the resource and the holding area associated with the resource. The queue and workstation are encapsulated within a container, making it easier to insert into a model. The resource router routes entities to the resources with the smallest E_D. The router does this by checking what resources the entity can use. Then computing the E_D of each resource the entity can use, determining the resource with the minimum E_D, and sends the entity to the resource with the smallest E_D.

FUTURE DIRECTIONS

Automated Input

Currently, NNI-OP uses Job Tracking Sheets for worker to use in the facility. The PPS tool requires all the information that is contained on a Job Tracking Sheet. NNI-OP would like the capability, of automate the collection of information from an already completed Job Tracking Sheet. This would save significant time for the planners, since they will not have to enter the same information twice. However this would require some additional information that is not present on the current Job Tracking Sheet.

Automated Output

NNI-OP envisions capability of an expert system to use output of the simulation as input for subsequent runs. The system will gradually make improvements based off the previous run output until user selected thresholds are reached. This requires modeling specific NNI-OP planning methodologies and creating rules for this expert system to follow.

Start From Non-idle System

NNI-OP would like to have the capability to capture the current state of the facility floor. This current state would include the status of current jobs and resources on the floor. The capability to capture the current state of the facility could be used to initialize the simulation from a non-empty and non-idle state. Starting from this current state would allow the simulation to produce more accurate output, given the current state of the facility.

"Hooks" left in system

The Senior Capstone Design Team (SCDT) is providing interfaces between the three main sub-systems of the PPS tool. These interfaces will allow NNS to interchange the input and output sub-systems and replace them. The only requirement is that the format of the interface must be kept intact in order to allow the PPS Simulation to run properly.

CONCLUSION

By having access to the PPS tool, NNI-OP will be able to visualize and interpret the impact that these new jobs will have on the system. The tool will allow the user to make better decisions on whether to accept or refuse a new job. The user has the ability to define any set of resources, thus effectively representing the floor of any desired facility. This feature gives the PPS tool great adaptability and reusability.

The SCDT has already submitted a proposal and a design that has already been accepted by NNI-OP. The SCDT has moved into the prototype development phase of the project. A prototype review is scheduled for early April, 2014. This prototype review will allow the SCDT to demonstrate the functionality that is present in the prototype, and it will give NNI-OP the opportunity to provide feedback. The SCDT will deliver a functional prototype to NNI-OP by the end of the spring semester, which is in May, 2014.

ACKNOWLEDGEMENTS

The authors would like to thank Newport News Industrial (NNI) for their participation and the valued input provided by James Whitley, Brian Bangs, and Erin Schiller. The authors would also like to thank Rob Lisle, Chris West, Irin Hall, John Lillard, Erick Hagstrom and the others from Newport News Shipbuilding (NNS), Department K76, for their support throughout the development process. The authors are especially grateful to Rex Wallen. He served as the subject matter expert and as the main interface between the SCDT, NNI, and NNS. Finally, the authors would like to give a very special thanks to Dr. Roland Mielke and Dr. James Leathrum, Jr. Without their support, expertise, and instruction, this would not have been possible.

REFERENCES

[1] A. Allen, J. Caldwell, D. Drake, C. Heard, I. Sakiotis. "Discrete Event Simulation for Supporting Production Planning and Scheduling in Job Shop Facilities." *ModSim World.* Hampton, VA, 2014.

[2] Kádár B, Pfeiffer A, Monostori L. "Discrete Event Simulation for Supporting Production Planning and Scheduling Decisions in Digital Factories". *37th CIRP Int. Seminar on Manufacturing Systems; Digital Enterprises, Production Networks*. Hungary, 2004; 444-448.

BIOGRAPHY

Jesse Caldwell is a senior in the Modeling and Simulation Engineering program at Old Dominion University. He is interested in serious gaming, formal methods, simulation architecture, and simulation applications. He is looking forward to working in the industry upon graduating in May 2014.

Christopher Heard is a senior in the Modeling and Simulation Engineering Program at Old Dominion University. He is planning to continue his education by pursuing a Master's degree while working full time. He would like to find a job focused on modeling and simulation in the field of meteorology.

Ashton Allen is a senior in the Modeling and Simulation Engineering program at Old Dominion University. He plans to continue his education by completing the accelerated Master of Engineering program in Modeling and Simulation. He is interested in serious gaming, discrete event simulation, and statistical methods. Upon graduating in May 2014, Ashton will be working at the Newport News Shipbuilding in the Modeling and Simulation department.

Daniel Drake is a senior in the Modeling and Simulation program at Old Dominion University. He has spent several years with the U.S Navy as a cryptologic technician before pursuing his education at Old Dominion. During his time at Old Dominion, Daniel has developed simulations with application in commerce, virology, biology, defense, and manufacturing.

Ioannis Sakiotis is a senior in the Modeling and Simulation Engineering program at Old Dominion University. He plans to pursue his education by pursuing a Master's degree in Modeling and Simulation. His interests include autonomous robotic systems, transportation, serious gaming, and military applications of modeling and simulation.

Khatera Alizada

Afghanistan and US bargaining over Bilateral Security Agreement

Abstract

This paper examines the ongoing negotiations and bargaining between the United States of America (US) and Afghanistan over Bilateral Security Agreement (BSA). Currently a game of chicken is going on between the US and Afghanistan. The findings of this paper show that both Afghanistan and the US would be better off reaching a deal. Even though both would be worse off if they do not reach a deal, the efforts by both sides to achieve their best outcomes lead to delay in reaching a deal. When the recipient (Karzai/Afghanistan) believes that he can gain more than what is being offered, he would not accept the offer even though he would be worse off when not reaching a deal. On the other hand, the other side (the United States) believes that the recipient would accept the offer when the offer is positive and makes the recipient better off. The misperception about the other actor's belief explains the delay in the bargaining process. The actors' beliefs about fairness of the offer and the availability of concessions play an important role in determining the outcome of bargaining.

This paper examines the ongoing negotiations and bargaining between the United States of America (US) and Afghanistan over Bilateral Security Agreement (BSA). Although an analysis of the bargaining situation suggests that negotiated agreements are feasible in which both sides benefit (and therefore a Rubinstein-type bargaining model implies that resolution will be reached), concerns by both sides about the fairness of negotiated outcomes, and an unwillingness to yield further concessions has generated a negotiating game of chicken in which deadlock and failure of the negotiations is quite likely.

In the literature review, the paper discusses different bargaining models, processes and identifies the model (s) that can be applied in this case study. Later it discusses the interests of both parties in reaching the deal. Then it looks at the bargaining processes between Afghanistan and US over the BSA. It constructs a formal game model of this case study through application of the game of chicken and presents the conclusions.

What will bargaining produce and when will bargaining fail? “Actors bargain when there are many outcomes that both are willing to accept over the alternative to an agreement and they disagree about which of those outcomes is best” (Morrow, 1994, p.112). Both Afghanistan and the US bargain over the BSA, but disagree on which outcome is best. In bargaining, each side has a reservation point or level that the minimum deal that they can be accept. The area between the reservation points of the two sides is called zone of agreement, which includes all bargaining points that the two parties prefer over reaching no agreement. If there is no zone of agreement, then

bargaining theory implies that the two sides cannot reach an agreement because they do not have bargaining outcomes that they prefer over no agreement.

Within the zone of agreement, however, reaching an agreement remains challenging. The Nash bargaining solution provides one normative standard for a 'good' agreement. The utility space plots the two sides' utility gain from each deal over the conflict zone. The horizon or frontier shows the limit of the possible bargains to the players and the deals on the frontier are pareto-optimal deals that each player prefers over a deal under the frontier. The Nash bargaining solution specifies a set of properties that a solution should possess and identifies the outcome that meets these conditions (Morrow, 1994, p.114). "This solution is the point which maximizes the product of the difference between each player's payoff and his status quo position" (Lockhart, 1979, p.5).

Although the Nash bargaining solution describes what outcomes bargaining will result or an "optimal arbitration scheme," the Nash bargaining solution has some substantial limitations. It always reaches an agreement. However, in real life situation, the bargaining can break down even when one was possible to reach. Even though both parties prefer the BSA over no agreement, it is still possible that the bargaining break down when Afghanistan and the US have disagreement over the best outcome.

The Nash bargaining solution does not describe the reason why the parties reach an agreement. It forsees what deals are reached (Morrow, 1994, p.145). It would be hard to apply the Nash bargaining solution in the bargaining between US and Afghanistan because utilities for the players are difficult to quantify. For example, it is

difficult to quantify national pride, Taliban take over and so on. However, it can be used to analyze whether the deal maximize both parties payoff.

Walton and Mckersie (as cited in Lockhart, 1979) identify intraparty disagreement process where there is disagreement within negotiating parties (p.13). In the case of Afghanistan one can observe intraparty disagreement process. If Afghanistan is assumed one party and the US is considered another party. There are some among President Karzai's (Karzai) cabinets and the Loya Jirga, who prefer the immediate signing of the agreement, but Karzai wants to delay it. Over all, these four processes do not focus much on international relations, but some aspects can put light to international relation bargaining.

Rubinstein bargaining model would be more applicable to this case study. Rubinstein bargaining model looks at the incentives of the sides and discusses what offers should the sides make "and when they should accept those offers" And it asks "how rational actors bargain" (Morrow, 1994, p.145). Rubinstein bargaining follows a sequential offer process in which the first player makes an offer, the second player either accepts or rejects the offer. If rejected, the second player makes a counter offer and it goes back and forth. Afghanistan and the US make sequential offers. The side in control of how offers made has advantage.

There is the concept of discount factor in repeated bargaining. The smaller discount factor means that the player puts high value on the present (Morrow. 1994, p.38). If the discount value is smaller, the player values the present more than the future and is more impatient. In this case player one has the advantage of being player one and making the first offer if both players have the same discounted factor because

player two needs to wait and get affected by the discounted factor. Taking advantage of this player one would offer less than what player 2 would ask in the second round. The larger the pressure to reach the deal quickly means a smaller discount factor for both players; then player one receives a bigger share. Since reaching the deal quickly is important, player one will take advantage and get a larger share by pressurizing player two. This game does not have other subgame-perfect equilibria. There is no real bargaining, however, because the first offer is always accepted and there are no sequential offers actually realized. This is the limitation of Rubinstein bargaining model and it does not depict real life (p.148-9).

Bargaining models focus on two main issues the distribution issue “who wins” and “who loses” and the efficiency of bargaining. Does bargaining process result in outcomes that make all parties better off? Do they use all the resources or fail (McCarty & Meirowitz, 2007, p.217)? If the players agree, they receive utility. However, if they do not reach an agreement, each player gets non-agreement value of outside option. There should be a feasible allocation that each player prefers over disagreement values. “Bargainers have to always contend with the possibility that an agreement will not be reached and they will be left with their outside options” (McCarty & Meirowitz, 2007, p.219). Bargainers who take risk can also have the possibility to get better deals because of their aggressive demands. Lower discount factor depicts higher risk aversion and increase disagreement. Bargainers who accept tougher risk are more likely to get greater share from a deal or the deal completely breaks up (McCarty & Meirowitz, 2007, p.219).

Bearce, Floros & Mckibben (2009) argue that states are more likely to enter bargaining when shadow of the future is long. “When the shadow of the future is short and following Fearon (1998), the bargaining problem is corresponding small, we are unlikely to observe much international bargaining because states do not anticipate that any agreement reached in the bargaining phase would be suitable in the subsequent enforcement phase” (2009, p.719). Afghanistan and the US entered bargaining because of maintaining long term relationship so the shadow of the future will be long if they reach a deal. There international cooperation framework identifies three phases of bargaining: pre-bargaining, bargaining and enforcement. Currently, Afghanistan and the US are in the bargaining phase.

Potential Gains from Bilateral Security Agreement

The Bilateral Security Agreement could provide benefits for both countries. The United States will continue its funding to Afghanistan’s security forces and economic development. Without these funds it is difficult for the Afghan government to fund its military. It will also encourage the international community’s commitment to the security and development of Afghanistan. It can give assurance to the US that its 12 years of efforts was ended with a responsible withdrawal and Afghanistan does not become a safe haven for the Al-Qaeda and its affiliates because of which the US and its allies went to Afghanistan in the first place. If Afghanistan becomes a safe haven to Al-Qaida and its affiliates it can again pose serious threats to the world (Weinbaum, 2013). The United States hopes that its presence will give some stability to Afghanistan and persuade the Taliban to agree to some form of political settlement to a stalemate conflict. Any form of conflict in Afghanistan after US withdrawal without a security

agreement can threaten the stability and security of the region specially a nuclear Pakistan.

US and Afghanistan's long term alliance can discourage and prevent the undue intervention of Afghanistan's neighbors. With the establishment of a security alliance between Afghanistan and the US, Afghanistan's neighbors will be more cautious to exert negative influences on Afghanistan. Afghanistan needs the support of positive alliance to strengthen its security forces to defend itself against internal and external threats. Both Afghanistan and its neighbors can gain from a more stable economic and political stable Afghanistan. A more stable Afghanistan can bring economic prosperity to the region through the construction of a gas pipeline from Turkmenistan to Pakistan and the rest of the world. It can also connect Pakistan's goods' market to Central Asia (Weinbaum, 2013).

Currently Afghanistan is suffering from a high level of corruption and lacks the necessary infrastructure to ensure stability and security (Miller, 2013, p.93). If The Taliban and the external spoilers can easily take advantage of the situation, causing further deterioration. There is a lack of strong, functioning and organized government opposition to hold the government accountable and ensure the necessary checks and balances.

The war in Afghanistan has been a long war for the US and has cost 2,274 American service members' lives. (New York Times, Nov 6, 2013). The US is also wary of Afghanistan turning into safe have to Al-Qaeda and its network (Miller, 2013, p.87).General John Allen, Commander of ISAF and US forces in Afghanistan recommended keeping 20,000 troops for the purposes of counter terrorism and training

after 2014. The US needs a long term presence in Afghanistan to combat Al-Qaeda and its affiliates. Miller identifies some other US interests in the region including stability of Pakistan and the security of its nuclear weapons, Iranian regional influence and drug trade, and the humanitarian crisis with the collapse of Afghanistan and the US withdrawal from Afghanistan.

Afghanistan's fragile economy is dependent on for foreign funding. A complete withdrawal of foreign troops raises concerns about international funding to Afghanistan for the purpose of economic development and strengthening its security forces. It might have negative effect on economic investments in Afghanistan. The local and foreign businesses would be hesitant to invest in an unstable economy. In the past decade Afghanistan experienced tremendous economic growth. Despite tremendous progress in the health sector, telecommunication, education, women initiatives and other areas, Afghanistan is not yet ready to let the international community abandon it as in the 1990s, which led to chaos and then became the safe haven for Al-Qaeda and its affiliates.

Afghanistan is a transitional democracy. It takes time to strengthen its democratic institutions and it would be difficult to promote these institutions in isolation. It needs to reform the government institutions and deal with the corruption issue. Women initiatives are one the main achievements of the past decade. Afghanistan's isolation can jeopardize the gains of the past 10 years.

The US complete withdrawal would not be in the best interest of the US. If the US leaves Afghanistan and political sphere changes to a friendly regime to Al-Qaeda and its affiliates, once again Afghanistan becomes the host of international terrorist

network; it can not only jeopardize peace and stability of the region, but the world. It becomes easier for the terrorist groups to have their training camps and a safe space for planning their terror activities to destabilize global peace.

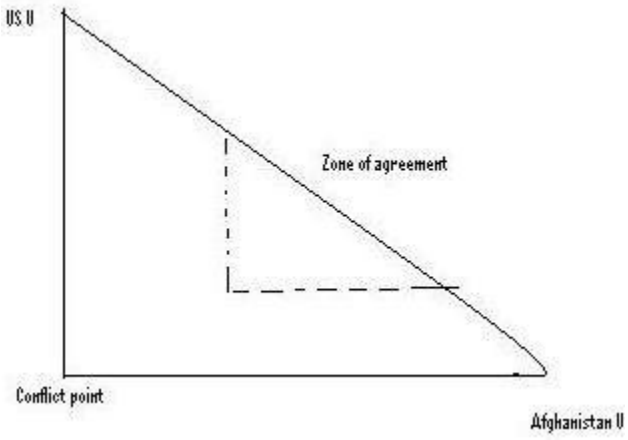
US withdrawal from Afghanistan can be tricky. Will the US be able to safely ship its cargos through Pakistan routes? The routes from Afghanistan to Pakistan port pass through areas under the influence of Taliban. In the past NATO shipments suffered from blockade of their shipments. Despite the contracts between Pakistan and US to use Pakistan routes until end of 2015, Pakistan government showed little effort to intervene the blockade when anti drone Pakistani protestors in Northwestern Khyber Pakhtunkhwa province halted NATO truck shipments from and to Afghanistan (Ali, Sahi, & Magnier, 2013). The protestors were mainly from Tehreek-e-Insaf party and the radical Islamist Jamaat-e-Islami party, both the ruling parties of the province. The Pentagon had to stop cargo shipments stating that the protests would put the lives of truck drivers and the contractors'' lives at risk.

If the utility of US payoff is higher in reaching this deal than withdrawing. US would make the concession to reach the deal. If the US payoff is lower in reaching this deal than making the concession, the US would withdraw without making the concessions. On the other hand if Karzai's payoff is higher in reaching the deal than his BATNA. Karzai would accept the agreement. If his payoff is lower in reaching the deal without the concessions, he would prefer the BATNA.

Figure 1 plots the utility gains of Afghanistan and of the US from reaching a deal over BATNA. Assuming that both parties prefer the deal over the conflict point and the bargains on the frontier are pareto-optimal outcomes and any deal under the frontier is

less efficient. Each would be worse off if they do not accept the agreement. But each side prefers distinct feasible bargains.

Figure 1: Utility gains of Afghanistan and the US



As negotiations developed over the BSA, four basic alternatives emerged. These are ranked below for each side.

Afghanistan:

1. No more raids on Afghan homes, help restart peace talks with the Taliban, release of Taliban prisoners from Guantanamo Bay prison
2. At least the other side meet some of the demands if not all
3. Agree to a deal without these concessions
4. BATNA (The conflict point, when no agreement (Morrow, 1994, p. 112)

USA:

1. Raids in extra ordinary circumstances when American lives are at risk, not guaranteeing peace with the Taliban, not releasing Taliban prisoners from Guantanamo Bay prison
2. Concession to at least some of the demands
3. Concession to all the demands

4. BATNA

Expected Bargaining Outcomes

It seems that a game of chicken is going on between Afghanistan and the US. A normal form game can be applied to give a different representation of Afghanistan US bilateral security negotiations. This model will portray the game of chicken (in which one actor can win if the other actor steers away from danger) to this case study assuming that both the US and Afghanistan receive equal payoffs.

The actual payoffs would depend on how much the US and Karzai value these options. In assigning the payoff value, the method used by Sokolowski and Banks (2009) in modeling Cuban Missile Crisis, is used (p.193). One can rank their preferences and assign a value for each choice from the least to most preferred option where 4 is the most preferred and 1 is the least preferred option. This method does not assign utility value since the utility for each option is unknown.

Table 1 illustrates a pay off matrix for US and Karzai. Karzai has two strategies make concession or do not make concession until US fulfills his demands. US also has two strategies to make concessions to Karzai's demands and not make concession hoping that Karzai would sign without US concessions. There are two Nash equilibria to this game. First Karzai does not make concession and US makes concessions. Second US does not make any concession and Karzai makes concession.

Table 1: Payoff Matrix: US and Karzai

Karzai/US	Concession	No concession
Concession	3,3	2,4
No Concession	4,2	1,1

The preferred strategy for each side in this game is that the other side meets his demands without him meeting the other side's demands. US prefers Karzai make concessions without the US making any concessions. Karzai prefers US making concession and he does not make any concession. The best winning outcome for each player is to hold firm to his demand and let the other chicken out (Morrow, 1994, p.93). However, both would be worse off if neither chicken out (if Karzai does not make concession and US does not make concession and they end up without any agreement).

However, one should be mindful of the payoff values that each actor receives from the options. It is hard to come up with these pay off values. If the payoffs are far from reality, this game would not be the representation of actual situation and the outcome or the result would be different.

Bargaining Over the BSA

The international community and the US repeatedly committed to continue their support for Afghanistan post 2014. The US committed to continue military support for counterterrorism operations, training Afghan security forces and economic developments. This support is conditioned to the signing of BSA (Felbab-Brown, 2013, p.65).

After negotiators had produced the BSA, the Loya Jirga (Grand Assembly) of 2,500 Afghan tribal elders and leaders endorsed the security agreement and recommended prompt signing of it. But bargaining soon broke down. Karzai rejected this recommendation and added a new set of demands. After Karzai refused to sign, the American officials believed he was bluffing (Craig & DeYoung, 2013). He demanded

that the American forces should immediately stop raids on Afghan homes, help with the peace process between Afghanistan government and the Taliban and he asked for assurance about the transparency in the elections. He threatened to break down the security agreement if US did not make concessions to these demands. Ending raids would give an end to American's last remaining combat mission. "In practice that would mean an end to the last remaining combat missions American troops are regularly carrying out: raids by elite units aimed at capturing high-profile insurgents" (Nordland, 2013, 1). Karzai also insists on delaying the signing of the agreement until after the presidential election in 2014. On the other hand, the Americans are asking for an immediate signing of the agreement to give the American and NATO enough time to plan the next phase after the combat missions terminates by the end of 2014. If the security agreement is reached, it would take effect January 1, 2015 and an estimated 10,000 US troops would be stationed across the country in US bases (Behn, 2013).

In response to Afghanistan's counter offer, the United States announced that Afghanistan should sign the deal by the end of 2013 or US would withdraw its troops completely. In a meeting with Karzai, Susan E. Rice, President Obama's top national security adviser told that US would withdraw all its troops from Afghanistan if Karzai failed to sign the agreement by the end of 2013 year. Not only Karzai did not sign the agreement, but he extended his set of demands to also include release of all 17 Taliban from Guantanamo Bay detention center in Cuba. "If Rice's unannounced visit to Afghanistan, her first solo trip abroad in office, was designed to convince Karzai that the Obama administration was not bluffing about a complete withdrawal, it did not appear to work" (Craig & DeYoung, 2013).

In their meeting Rice stressed that the US would not interfere in the election. Rice also told that failure to sign the agreement by the end of 2013 would jeopardize international pledge to fund 4 billion to assist the Afghan military and the 4 billion funding for the economic development of Afghanistan (Craig & DeYoung, 2013).

In this case the shadow of the future needs to be considered: whether shadow of the future is long or short. Karzai knows that the Americans will eventually leave whether 2 years or ten years. However, the Taliban will be there forever. Therefore, who would Karzai be siding with or making a deal with. Karzai's insisting on reaching an agreement and integration of the Taliban. Karzai seems inclined to be making an effort to be putting a deal with the Taliban by adding Taliban terms as preconditions to the deal (Engel, 2013).

The US claims that Karzai's demands are outside US zone of agreement. The conflicting point for both Karzai and the US is not reaching an agreement and their Best alternative to the negotiated agreement (BATNA). The claim of each side that the other side's demands are outside their zone of agreement is causing the delay to reach in reaching the deal. The US insists on rapid signature, but Karzai insists on delaying. This makes the US seem less patient and Karzai seems more patient by delaying to sign the agreement. Karzai by delaying is signaling strength of his position. He does not want to look weak by accepting the offer right away. As accepting the offer immediately would signal his weakness. On the other hand, some commentators in the US stated that it is Afghanistan who needs the US more than the US needs Afghanistan. These comments suggest that according to these commentators Afghanistan has a weaker bargaining

position and there is asymmetry in the negotiations. They want to signal that the US has bargaining power.

Karzai claims that it is more efficient to delay signing the agreement. While the US claims that it is more efficient to reach an agreement rapidly. When Karzai refused to sign the agreement without US meeting the new conditions, the US attempted to work around Karzai and suggested that one of his cabinet members, minister of defense, can sign the agreement. However, Karzai's spokesperson contested stating that the president would not authorize any of his cabinet members to sign the agreement and added that it cannot be signed until government demands are fulfilled. Karzai's spokesperson, Aimal Faizi said, "We are certain that the US can meet our conditions in practical terms within days or weeks" (Gearan & Craig, December 4, 2013).

A number of American senior officials met with Karzai after he refused to sign the agreement. Susan E. Rice, Obama's top security advisor, and James F. Dobbins Special Representative for Afghanistan and Pakistan met with Karzai to see if Karzai backed down on any of his demands, but he was still insisting on his conditions before signing the agreement. Dobbins told that the US is supporting peace process and he is willing to talk with the members of Afghanistan High Peace Council on how to initiate talks with the Taliban, but US cannot guaranty peace because the issue depends on the Taliban and their goals (BBC Persian, 6 Dec 2013).

A Democrat senator Carl Levin, the Chairman of Senate Armed Services Committee, suggested to the White House to ignore Karzai's demands and wait until the new president takes office in Afghanistan assuming that the new president would sign

the agreement, US should plan for keeping forces after 2014. He states that if US keeps insisting on quick signing or signing of the deal by one of Karzai's ministers, it will add to Karzai's mistaken perception that US needs Afghanistan more than Afghanistan needs the US (Londoño, December 7, 2013).

In one of his speeches last March prior to a planned press conference with the Hagel's first trip as secretary of defense, Karzai indirectly signaled as if the US was not serious in helping with the peace process. "The Afghan president delivered a speech in which he implied that the United States was colluding with the Taliban in order to justify fighting in his country longer" (Londoño, December 7, 2013). His remarks led to the cancelation of the press conference. It seems that Karzai believes the US is not serious in their attempt to peace processes with the Taliban to justify their long term war in Afghanistan. On the other hand, the US might have some mistrust about Karzai's long term commitment of Karzai to the agreed terms as Karzai's refusal to signing the agreement was unexpected.

Conclusion

This research shows the strategic alliance between Afghanistan and US is making both players better off, but also illustrates the challenges of maintaining that alliance. If BATNA is their least preferred option, one might expect an agreement to follow. But the breakdown of bargaining between the U.S. and Karzai illustrates the significant challenges of reaching agreement even in the face of potential mutual gains.

The analysis of the mixed strategy equilibrium shows that reaching the deal may be dependent on how much both actors prefer reaching a deal over their BATNA. Afghanistan needs this agreement for strategic reasons and the US needs for ensuring

security of the region, its goal of combating global terrorism and for a responsible end to the war.

However, when multiple bargains are possible, bluffing by both sides can create the potential for negotiation breakdown. And little consideration is given to asymmetry. If one of the actors has the bargaining power and is not offering a good enough deal, assuming that the other actor would accept it, it turns out that the other actor does not accept. It leads to a delay in reaching an agreement.

When the recipient (Karzai/Afghanistan) believes that he can gain more than what is being offered, he would not accept the offer even though he would be worse off when not reaching a deal. On the other hand, the other side (the United States) believes that the recipient would accept the offer when the offer is positive and makes the recipient better off. The misperception about the other actor's belief explains the delay in the bargaining process. The actors' beliefs about fairness of the offer and the availability of concessions play an important role in determining the outcome of bargaining.

Bibliography

Ali, Z., Sahi, A. & Magnier, M. (2013, December 4). U.S. halts truck shipments through Pakistan amid anti-drone protests. *LA Times*. Retrieved from http://www.latimes.com/world/worldnow/la-fg-wn-pakistan-afghanistan-nato-truck-blockade-20131204,0,2693528.story#ixzz2mqHOIzbB

Associated Press. (2013, December 3). US stops shipping from Afghanistan during protests. *Washington Post.* Retrieved from http://www.washingtonpost.com/politics/us-stops-shipping-from-afghanistan-during-protests/2013/12/03/c816b3fe-5c69-11e3-8d24-31c016b976b2_story.html

BBC Persian. (2013, December 1). *BBC*. Retrieved from http://www.bbc.co.uk/persian/afghanistan/2013/12/131201_zs_afghan_security_council_us_ansf.shtml

BBC Persian. (2013, December 2). *BBC Persian*. Retrieved from http://www.bbc.co.uk/persian/afghanistan/2013/12/131202_k02-coalition-ansf-fuel.shtml

BBC Persian. (2013, December 1). *BBC Persian*. Retrieved from http://www.bbc.co.uk/persian/afghanistan/2013/12/131201_zs_afghan_security_council_us_ansf.shtml

BBC Persian. (2013, December 6). Retrieved from http://www.bbc.co.uk/persian/afghanistan/2013/12/131206_dobbins_bsa_karzai.shtml

Bearce, D. H., Floros, K. M., & Mckibben, H. E. (2009). The shadow of the future and international bargaining: The occurrence of bargaining in a three-phase cooperation framework. *The Journal of Politics*, 71(2), 719–732.

Behn, S. (2013, November 26). US, Afghan tensions rise over security agreement. *VOA News.* Retrieved from http://www.voanews.com/content/us-afghan-tensions-rise-over-security-deal/1797911.html

Biddle, S. (2013). Ending the war in Afghanistan: How to avoid failure on the installment plan. *Foreign Affairs*. September/October. Retrieved from http://www.foreignaffairs.com/articles/139644/stephen-biddle/ending-the-war-in-afghanistan

CBS News. (2013, October 3). Afghanistan: U.S. blocking deal on future security pact. *CBS News*. Retrieved from http://www.cbsnews.com/news/afghanistan-us-blocking-deal-on-future-security-pact/

Craig, T. & DeYoung, K. (2013, November 25). Karzai tells Susan Rice of more demands for accord extending U.S. troop presence. *Washington Post.* Retrieved from http://www.washingtonpost.com/world/national-security-adviser-susan-rice-visits-afghanistan-amid-tension-over-troop-accord/2013/11/25/fd0f8460-55dd-11e3-835d-e7173847c7cc_story.html

Craig, T. (2013, December 1). Karzai government accuses U.S. of withholding fuel from Afghan forces. *Washington Post.* Retrieved from http://www.washingtonpost.com/world/karzai-government-accuses-us-of-

withholding-fuel-from-afghan-troops/2013/12/01/74b2260a-5ac3-11e3-801f-1f90bf692c9b_story.html
Dna. (2013, December 3). Afghan-US bilateral security agreement rattles Pakistan. *Dna.* Retrieved from,http://www.dnaindia.com/world/report-afghan-us-bilateral-security-agreement-rattles-pakistan-1928939
Engel, R. (2013, November 29). How would Afghanistan look without the United States' support? *NBC News. Retrieved from* http://worldnews.nbcnews.com/_news/2013/11/29/21676431-how-would-afghanistan-look-without-the-united-states-support
Felbab-Brown, V. (2013, Fall). Afghan after ISAF prospects for Afghan peace and security. *Harvard International Review*. 65-69.
Gearan, A. & Craig, T. (2013, December 4). U.S. looks for work-around to Afghan security impasse. *Washington Post.* Retrieved from http://www.washingtonpost.com/world/national-security/us-looks-for-work-around-to-afghan-security-impasse/2013/12/04/d250c6ae-5cfd-11e3-95c2-13623eb2b0e1_story.html
John Kerry Secretary of State. (2013, November 24). The Loya Jirga and the U.S.-Afghanistan Bilateral Security Agreement [Press Statement]. Retrieved from http://www.state.gov/secretary/remarks/2013/11/218031.htm
Lockhart, C. (1979).*Bargaining in international conflicts*. New York: Columbia University Press.
Londoño, E. (2013, December 7). Hagel in Kabul for unannounced visit, but no plans to meet with Karzai. *Washington Post.* Retrieved from http://www.washingtonpost.com/world/hagel-in-kabul-for-unannounced-visit-but-no-plans-to-meet-with-karzai/2013/12/07/ab69eb8a-5f30-11e3-bc56-c6ca94801fac_story.html
McCarty, N., & Meirowitz, A. (2007). *Political game theory: An introduction*. Cambridge: Cambridge University Press.
Michaels, J. (2013, December 7). U.S. and Afghans plan for future despite uncertainty *USA Today* Retrieved from http://www.usatoday.com/story/news/world/2013/12/06/karzai-afghanistan-coalition-security-army-pentagon/3895847/
Miller, P. D. (2013, January 31). The US and Afghanistan after 2014. *Survival: Global Politics and Strategy,* 55(1), 87-102, DOI: 10.1080/00396338.2013.767406
Morrow, J. D. (1994). *Game theory for political scientists*. Princeton: Princeton University Press.
New York Times. (2013, November 6). U.S. military deaths in Afghanistan. *New York Times*. Retrieved from http://www.nytimes.com/2013/11/07/us/us-military-deaths-in-afghanistan.html?_r=0
Nordland, R. (2013, November 24). Elders back security pact That Karzai won't sign *New York Times*. Retrieved from http://www.nytimes.com/2013/11/25/world/asia/afghan-council-approves-us-security-pact.html?pagewanted=2&hp
Nordland, R. (2013, November 30). Afghans assail Karzai's disparate views on killings.

New York Times. Retrieved from http://www.nytimes.com/2013/12/01/world/asia/another-afghan-child-dead-a-different-response.html?hpw&rref=world&_r=0

Nordland, R. & Masood, S. (2013, November 29). Recent drone strikes strain U.S. ties with Afghanistan and Pakistan Retrieved from http://www.nytimes.com/2013/11/30/world/asia/drone-strike-pakistan.html?_r=0

Nordland, R. Rubin, A.J. (2013, November 26). Karzai's bet: U.S. is bluffing in warning on security pact. *New York Times*. Retrieved from http://www.nytimes.com/2013/11/27/world/middleeast/karzais-bet-us-is-bluffing-on-warning-on-security-pact.html?src=recg

Peterson, S. (1996). *Crisis bargaining and the state: The domestic politics of international conflict*. Ann Arbor: University of Michigan Press.

Sokolowski, J. A. & Banks, C. M.(2009). *Modeling and simulation for analyzing global events. Hoboken*: Wiley and Sons,

Shanker,T & Ahmed, A. (2013, December 7). In Afghanistan, Hagel presses for pact on security, but is not meeting Karzai. Retrieved from http://www.nytimes.com/2013/12/08/world/asia/hagel-in-afghanistan.html?_r=0

Weinbaum, M. (2013, November 21). The tortuous route of the U.S.-Afghan security pact. *Middle East Institute*. Retrieved from http://www.mei.edu/content/tortuous-route-us-afghan-security-pact

Water Security in the Kabul-Kunar River Basin

Amanda Norton
Old Dominion University
anort009@odu.edu

Keywords: Water security, Afghanistan, system dynamics

Abstract

Water security in Afghanistan is a global concern that influences the political and economic stability of the region. This paper uses a system dynamics model to explore four factors that influence the water security in Afghanistan's Kabul-Kunar River Basin. The water security is modeled using the dependent variables of hydroelectric capability, stored water capacity, and arable land acreage. A literature review revealed four influencing factors: transborder politics, foreign investment, environment, and economic growth, which were converted to indices that act on the dependent variables. The paper presents a model that can be used to predict the future of water security in the region.

1. INTRODUCTION

Water security is a global concern and researchers are looking toward Afghanistan in order to address one potential source of instability. Availability and utilization of water not only impacts a country's internal economy and population, but also influences regional, transborder politics and economics. Additionally, as stated by former United Nations Secretary-General Kofi Annan, "fierce competition for fresh water may well become a source of conflict and wars in the future" [Gonzalez 2013c]. Improving water security in volatile areas can result in greater regional stability, which has a positive ripple effect worldwide.

This report details the research question and methodology behind modeling water security in the Kabul-Kunar River Basin in Afghanistan. Originally the goal was to model water security in all of Afghanistan, but due to the complexity of that project, the scope was narrowed to the Kabul-Kunar River Basin, which is the most densely populated of the five river basins and encompasses the country's capital city [Favre and Golam 2004]. This report will provide a background on water security and the Kabul-Kunar River Basin, describe the research question, discuss the development of the model, provide analysis of the model, and conclude with suggestions for future projects.

2. BACKGROUND

Afghanistan has access to an abundance of existing natural water sources [Gonzalez 2013c]. The Kabul-Kunar River Basin alone contains ten river systems with 22 billion cubic meters of available water [Qureshi 2002]. The river basin flows through or along the borders of eleven provinces, serving over seven million people, and accounting for 26 percent of the annual water flow in Afghanistan [GIRoA 2007]. Additionally, the river basin supports over 300,000 hectares of intensively irrigated land [GIRoA 2007].

Despite enough natural water to support its population, Afghanistan's water security remains unstable due to a lack of infrastructure to store, distribute, and supply the water for agricultural and human consumption [Gonzalez 2013c]. Additional problems stem from a lack of transborder political agreements regarding the development of infrastructure that may impact regional countries' water security. Pakistan relies on water originating in Afghanistan for agriculture, energy, and water used by its population [GIRoA 2007]. Pakistan has raised concerns over dam construction projects in Afghanistan, citing a 16 to 17 percent drop in water supply to their nation [INPAPERMAGZINE 2011]. Even though both countries have an interest in water management and security, no treaties currently exist between Afghanistan and Pakistan.

3. RESEARCH QUESTION

The research question was developed by first analyzing the required task and desired assessment. The task was developed by determining what variables influence water security. The three variables chosen are arable land area measured in hectares (ha), hydroelectric power production capability measured in megawatts (MW), and water storage capacity measured in millions of cubic meters (Mm^3). After a review of the literature, it was determined that the variables were most influenced by transborder politics, foreign investment, environmental changes, and local economic growth.

Thus, the task was defined to characterize the changes in arable land area, hydroelectric power production capability, and storage capacity of water in the Kabul-Kunar River Basin based on changes in transborder politics, foreign investment, the environment, and local economic growth. The goal is then to assess the effects or outcomes of executing transborder political agreements and changing foreign investment. Transborder political agreements and foreign investment have the most impact on Afghanistan's water security. The environment is comparatively constant and unchanging, and economic growth is greatly influenced by the water security. Only after improving water security

through transborder political agreements and foreign investment can economic growth be reevaluated and the changes incorporated into the model.

This task and assessment produces the following research question: How can transborder political agreements, foreign investment, the environment, and local economic growth changes be measured, then be represented qualitatively in a developed and quantifiably supported model that can predict changes in arable land acreage, hydroelectric power production capability, and water storage capacity in the Kabul-Kunar River Basin in Afghanistan?

4. MODEL DEVELOPMENT

The first step of model development is to choose a modeling paradigm. Systems dynamics was chosen as it allows for a top-down view of the system changing over time. Since this project is looking at a large, regional view of water security, systems dynamics proved to be the most appropriate paradigm. Next, a review of the literature was conducted and dependent variables were identified. As discussed in the previous section, the dependent variables are hydroelectric capability (MW), stored water capacity (Mm^3) and arable land (ha).

The literature review also identified four indices that have the most influence on the dependent variables. The four indices are described below:

Transborder Political Index: a measure of political influences from neighboring countries regarding water security. For the Kabul-Kunar River Basin, the primary influencer is Pakistan.

Foreign Investment Index: a measure of monetary foreign investment provided to improve water infrastructure in Afghanistan.

Environmental Index: a measure of environmental factors that impact water security, including natural factors such as rainfall and glacier melt, and man-made factors such as pollution.

Economic Growth Index: a measure of the impact of water security on the local economy.

The indices provide a convenient method of converting qualitative data into quantitative values that can be used in the model. During the literature review, key points and concepts were extracted for each index and assigned a number using a 1-5 Likert Scale. A positive value indicates a positive contribution toward improving water security in the Kabul-Kunar River Basin and a negative value indicates something that detracts from the area's water security. The literature and indices encompass a time period from 2001 to 2013. Appendix A provides tables for each index. The values of each entry in the index tables are summed to provide a composite score, as shown in Table 1.

Table 1. Index Composite Scores

Index Name	Composite Score
Transborder Political Index	-28
Foreign Investment Index	69
Environmental Index	-11
Economic Growth Index	-5

The next step in model development is creating the causal loop diagram. The causal loop diagram, shown in Appendix B, depicts the feedback relationships between variables. The dependent variables are shown in the boxes and the independent variables are shown as plain text. The arrows are color-coded to make viewing the diagram easier. The Foreign Investment Index loops are shown in orange, Transborder Political Index loops are light blue, Environmental Index loops in dark red, and Economic Growth Index loops are green. The blue loops show interactions among the other variables, and the pink arrows highlight negative feedback loops.

The negative feedback loops are defined through common sense logic of how the variables interact. An increase in hydroelectric capability results in a decrease in stored water because more of the water is used to generate electricity. An increase in arable land also results in a decrease in stored water as more water is being used for irrigation. Finally, an increase in arable land results in a decrease in aquifer projects because the arable land consumes the land that could be used to develop aquifers. The Environmental Index has a negative relationship to Sanitation System Projects because as the environment positively influences water security it decreases the need for sanitation projects.

Next, a stock and flow diagram was created in order to implement the model. The stock and flow diagram is given in Appendix C. The stock and flow diagram is similar to the causal loop diagram with the only major difference being the introduction of flow variables. The colors were chosen arbitrarily to improve visual understanding of the model. Constant variables are displayed given brown text. The Transborder Political Index loops are colored light blue, the Foreign Investment Index loops are green, the Environmental Index loop is dark red, and the Economic Growth Index loops are orange. The negative feedback loop between water storage and hydroelectric capability is shown in bright pink, the negative feedback loop between water storage and arable land is shown in red, and the negative feedback loop between arable land and aquifer projects is shown in light pink. All arrows are assumed to have a positive relationship unless otherwise indicated with a negative (-) sign at the arrow head.

The constant values assigned to the indices are their composite scores listed in Table 1 above. Population Growth is assigned a constant value of 1.56, relating to the average percent increase of the population in the provinces located within the Kabul-Kunar River Basin [The World Bank 2010]. The initial values of the dependent variables come from the literature. The initial value of arable land is 292,980 ha rounded to 300,000 ha [The World Bank 2010], storage capacity is 3309 Mm^3 [The World Bank 2010], and hydroelectric capability is 281 MW [Ahmad 2010].

The equations in the stock and flow diagram are found by summing the input variables. In the cases where the dependent variables are inputs into other variable functions they are scaled by dividing by their initial values. A list of equations used in the model is given in Appendix D. This methodology was chosen to provide a means for the model to execute conceptually. Due to a lack of data depicting the change in the dependent variables over a specified time period, the model is not calibrated. The data given in the literature review generally discusses future projects and gives the expected change in dependent variable values. While this data could be used to generate more accurate equations for the Dam, Sanitation, Aquifer, and Irrigation Project variables, it would detract from the purpose of the model, which is to demonstrate how the indices influence the dependent variables.

The model is programmed to run for a period of ten years with a time step of half a year, or six months. This allows adequate time to generate results. The results of the model are discussed in the next section.

5. ANALYSIS

Prior to executing the model, some expected analysis was conducted. Logically, increasing the Transborder Political Index will increase dam and irrigation system production and improve water security. Additionally, increasing the Foreign Investment Index and Economic Growth Index will increase all production efforts and improve water security. Finally, increasing the Population Growth will deplete the stored water capacity and the water storage development rate will need to increase in order to compensate for this.

Running the model using the derived equations and initial values defined in the above section, the following results are given in the figures below.

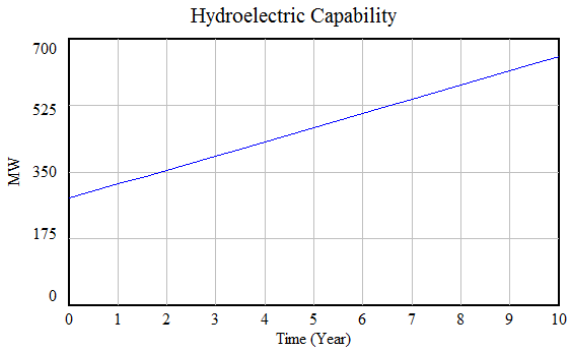


Figure 1. Hydroelectric Capability

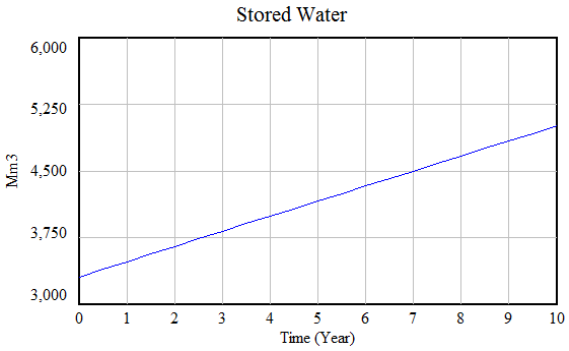


Figure 2. Stored Water

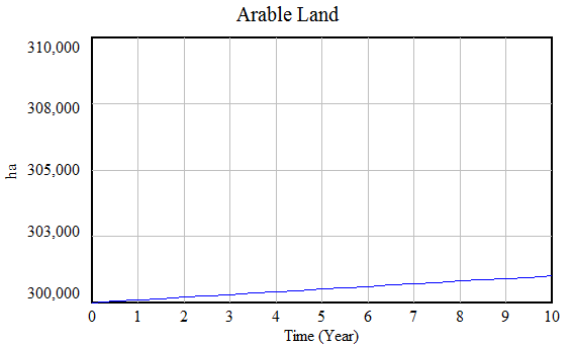


Figure 3. Arable Land

As depicted in Figures 1 -3, the infrastructure values increase linearly over the course of ten years. Arguably, this is because the Foreign Investment Index is great enough to counteract the negative influence of the other indices. In order to determine if the model reacts as expected, the index values can be arbitrarily changed in order to observe the behavior of the output. For this trial run, the Transborder Political Index was increased to 10, the Economic Growth Index changed to 20, and the Foreign Investment Index changed to 100.

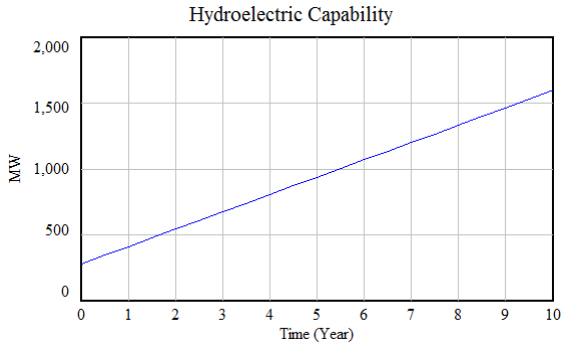


Figure 4. Hydroelectric Capability

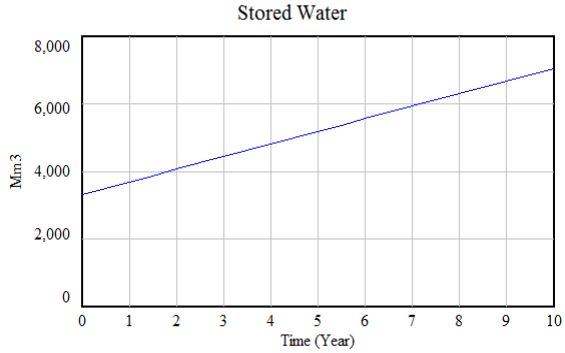


Figure 5. Stored Water

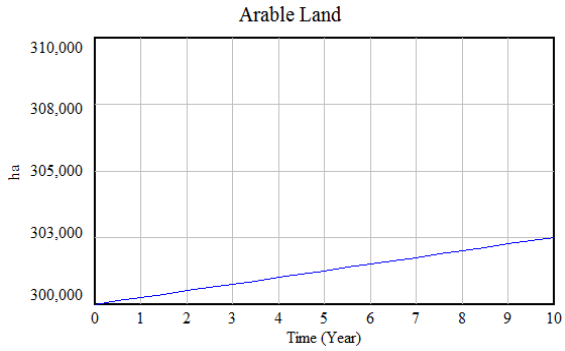


Figure 6. Arable Land

Figures 4 –6 show a significant increase in infrastructure development after changing the index values. This demonstrates that the model executes as expected. Once data is available to calibrate the model, the model can then be used to show how changes in the various index values impact each of the dependent variables, thus impacting water security in the Kabul-Kunar River Basin.

6. CONCLUSION

The purpose of this project was to develop a model that measures transborder political agreements, foreign investment, the environment, and local economic growth changes in a quantitative manner and then used to predict changes in arable land area, hydroelectric power production capability and water storage capacity in the Kabul-Kunar River Basin in Afghanistan. This allows users of the model to analyze and predict the future of water security in the region. In its current state, the model provides a conceptual look at how the variables interact. By incorporating calibration data, it would be possible to use the model to simulate the outcomes of actual infrastructure development based on changes in politics, economics, investment, and the environment.

References

Ahmad, Shahid, 2010, *Towards Kabul Water Treaty: Managing Shared Water Resources*. Rep. IUCN: 2. http://cmsdata.iucn.org/downloads/pk_ulr_d3_1.pdf

Availability of Water in the Kabul Basin, Afghanistan. 2010-3037, May 2010. USGS. http://pubs.usgs.gov/fs/2010/3037/pdf/fs2010-3037.pdf

Favre, Raphy, and Golam Kamal, 2004, *Watershed Atlas of Afghanistan Volume II Part III*. AIZON: 65. http://aizon.org/ws_volume%20IandII.htm

Gonzalez, Rainer, 2012, "The Decision to Plant Poppies: Irrigation, Profits, and Alternatives Crops in Afghanistan." *Afghanistan in Transition*. Civil-Military Fusion Centre. https://www.cimicweb.org/cmo/afg/Documents/Social_Infrastructure/CFC_Afghanistan_Poppies-Profits-and-Irrigation_Aug2012.pdf

Gonzalez, Rainer, 2013, "Social and Strategic Infrastructure." *Afghanistan Review*. Civil-Military Fusion Centre. https://www.cimicweb.org/cmo/afg/Pages/AFGReview.aspx

Gonzalez, Rainer, 2013, "Water Security: Afghanistan Transboundary Water Resources in Regional Context." *Transboundary Issues*. Civil-Military Fusion Centre: 4-5 https://www.cimicweb.org/cmo/afg/Documents/Social_Infrastructure/201310_CFC_Afghanistan_Transboundary_Resources_final.pdf

Government of the Islamic Republic of Afghanistan (GIRoA), 2007, *Transboundary Water Issues*, Appendix: 7. http://gfipps.tamu.edu/Afghanistan/Transboundary%20Water%20Issues26-04-07.pdf

Granit, Jakob, Anders Jagerskog, Rebecca Lofgren, Andy Bullock, George De Gooijer, Stuart Pettigrew, and Andreas Lindstrom, 2010, *Regional Water Intelligence Report Central Asia*. Rep. no. 15. Water Governance Facility. http://www.watergovernance.org/documents/WGF/Reports/Paper-15_RWIR_Aral_Sea.pdf

Mustafa, Khalid, 2011, "India to Help Afghanistan Build 12 Dams on Kabul River." *The International News*. http://www.thenews.com.pk/TodaysPrintDetail.aspx?ID=5933&Cat=13&dt=5/12/2011

Qureshi, Asad, 2002, *Water Resources Management in Afghanistan: The Issues and Options*. Working paper no. 49. International Water Management Institute: 7. http://www.afghaneic.net/library/hydrological%20surveys/wor49.pdf

"Rehabilitating Irrigation in Afghanistan." *What We Do: Focus on Afghanistan*. FAO Water Unit. http://www.fao.org/nr/water/news/afghanistan.html

S. Rep. No. 112th Congress-112-10, 2011, "Avoiding Water Wars: Water Scarcity and Central Asia's Growing Importance for Stability in Afghanistan and Pakistan." http://www.foreign.senate.gov/press/chair/release/?id=0b32e452-9c4c-4417-82ee-d201bcefc8ae

"Sharing Water Resources with Afghanistan," 2011, *Dawn.com*. INPAPERMAGZINE. http://beta.dawn.com/news/673055/sharing-water-resources-with-afghanistan

United States. Government Accountability Office. GAO-11-138, 2010, *Afghanistan Development: US Efforts to Support Afghan Water Sector Increasing, but Improvements Needed in Planning and Coordination*. http://www.gao.gov/products/GAO-11-138

The World Bank, 2010, *Scoping Strategic Options for Development of the Kabul River Basin*, Rep. no. 52211: 12, 17. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2010/04/12/000333037_20100412001029/Rendered/PDF/522110ESW0Whit1anistan0Final0Report.pdf

Biography

Amanda Norton is a graduate student at Old Dominion University. She plans to graduate in May 2014 with a Master of Engineering in Modeling and Simulation. She is a University of Colorado alumna, receiving a Bachelor of Science in Applied Mathematics with a minor in Computer Science in 2008. Amanda is a Lieutenant in the United States Navy and has completed several overseas deployments including a tour in Kabul, Afghanistan where she served as the lead convoy commander and intelligence, operations, and communications officer for the Counterinsurgency Training Center - Afghanistan. She is currently stationed at the Center for Surface Combat Systems Detachment East in Norfolk, Virginia.

Appendix A
Index Tables

Transborder Political Index	
Factors	**2001-2013**
Water Security Risk Index: High Risk	-4
No existing bilateral treaty	-5
Pakistan and Iran officials raise concerns that infrastructure developments would negatively affect their countries' water security	-5
Ecological, economic, political impacts downstream of irrigation projects have not been fully considered	-5
Concerns in Pakistan arose after announcement that India would provide assistance in building 12 dams	-5
Only applicable treaty was signed between Afghanistan and British in 1921; neither Afghanistan nor Pakistan recognize the current validity of this treaty	-3
Lack of essential hydrological data and technological capacity makes it difficult to monitor and implement/enforce any water-distribution agreements	-3
The Kunar River originates in the Pakistani Hindu Kush in Chitral District and crosses the border into Afghanistan to become a tributary of the Kabul River, which subsequently enters Pakistan as part of the Indus River Basin. In order to avoid greater losses of water through the Kabul River, Pakistan has already diverted flow of the Kunar River before it crosses Afghan border	3
Afghanistan and Pakistan unable to reach a water management/sharing agreement after several attempts	-3
The World Bank is currently leading an initiative to establish an agreement	5
Afghan, Pakistani and international experts met in a conference on "Regional Water Governance: Facing Scarcity, Enhancing Cooperation" in October 2012	3
The World Bank is aiming to enhance institutional capacity in order to improve the sharing of hydro-meteorological data	3
Since the Kabul River originates in Afghanistan it is able to use water control upstream to wield more power against its neighbors	5
Afghanistan is withdrawing less water from transboundary rivers than it would have legally been allocated by international agreements (about 30%)	3
Even though Afghanistan is not making full use of transboundary water resources, the ecosystems of the lower riparian states are already experiencing significant pressures.	-3
Afghanistan and Pakistan are moving towards joint management of shared basins starting with the construction of a 1500 MW hydropower projection the Kunar River	5
Afghanistan's initiative for construction of multi-purpose water projects on the tributaries of Kabul River would adversely impact Pakistan	-3
Pakistan is estimated to supper 16-17% drop in water supply from Afghanistan after construction of 13 dams on the Kabul River	-5
Various canals are developed off the Kabul river to irrigate Peshawar Valley and have greatly contributed to the prosperity of Charsadda District	3
Afghanistan lacks sufficient dams, reservoirs and flow control structures to adequately manage and control this runoff. As a result, the country has little control of water flow into neighbouring countries.	3
Pakistan is one of the most water-stressed countries, a situation likely to worsen into outright water scarcity owing to high population growth	-3
Pakistan is dependent on a single river system and lacks the robustness that many countries enjoy by virtue of having a multiplicity of river basins and diversity of water resources.	-5
Even, under the Indus Water Treaty, Pakistan is supposed to receive 55,000 cusecs of water, but authorities complain that its share was drastically reduced, causing damage to crops.	-3
Pakistan received 13,000 cusecs during winter and a maximum of 29,000 cusecs during summer.	-5
The prime minister of Pakistan has already established Pakistan Transborder Water Organization (PTWO) to tackle issues arising from construction of dams and water sector projects by upper riparian countries.	3
Afghanistan is a member of the Central Asia Regional Economic Cooperation Program (CAREC)	2
Afghanistan is a member of the Economic Cooperation Organization (ECO)	2
Overseas Development Assistance reached the equivalent of 83 USD per person in Afghanistan in 2008	3
The US pledged to donate 10.4 billion USD of development aid to Afghanistan between 2002 and 2008	3
Only 5 of 10.4 billion pledged has actually been dispersed	-3
Pakistan has significant increased its water use of the Indus River for power, municipal and agriculture during the last 30 years, and has higher water demand than can be currently met	-5
Pakistan is benefiting from flows from Afghanistan but provides no financial support for flow control structures or management of the river within Afghanistan	-3
Monitoring stations should be established to measure flows into Pakistan from Kabul and other rivers	-3
COMPOSITE SCORE	**-28**

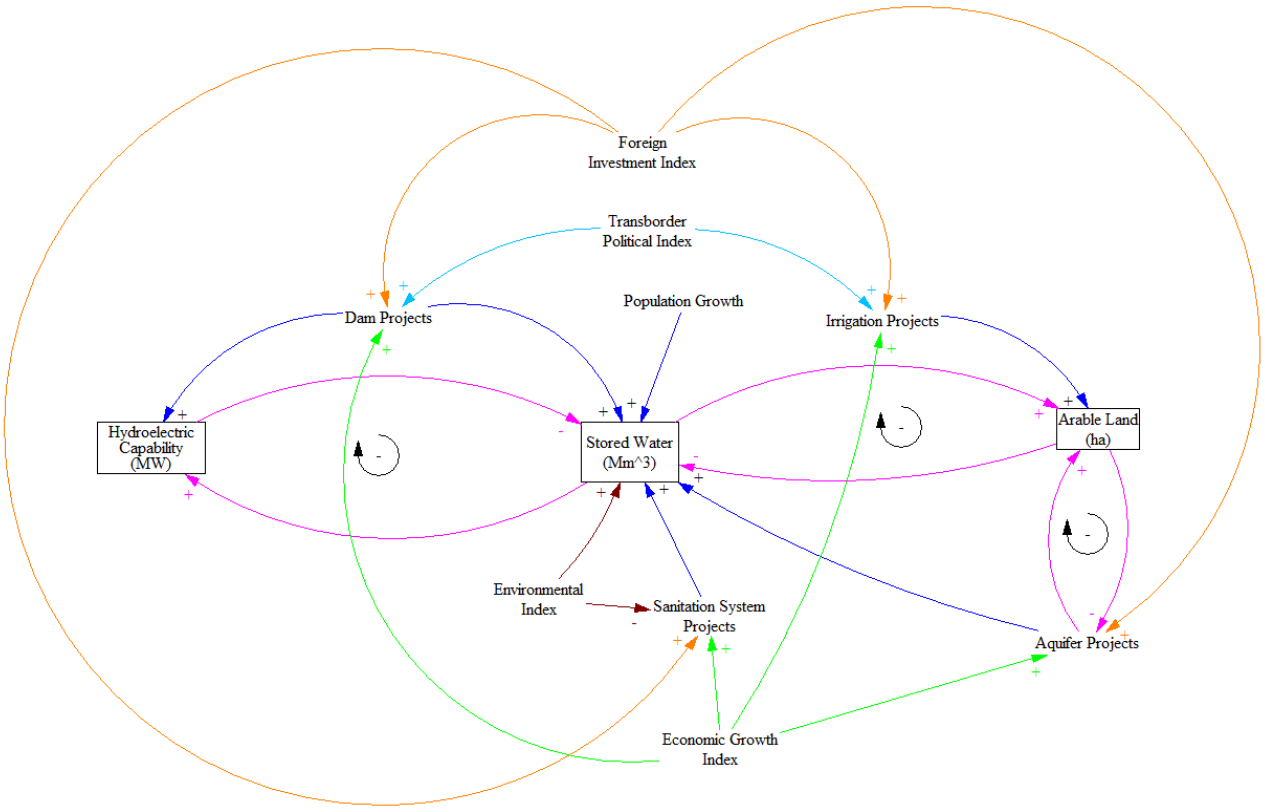
Foreign Investment Index	
Factors	**2001-2013**
1 USD in water investment results in between 3-34 USD in growth	3
International community has funded several irrigation projects across Afghanistan	5
India to contribute USD 6.8 billion toward construction of 12 dams	3
World Bank to contribute USD 8 million toward developing hydrologic, hydraulic and economic models	3
USD 332 million contributed toward Totumdara hydropower project	4
USD 1.174 billion Barak hydropower project	5
USD 1.078 billion Punjshir hydropower project	5
USD 72 million Haijana dam project	3
USD 207 million Kajab dam project	4
USD 356 million Tangi Wadag dam project	4
USD 51m Gat dam project	3
USD 442 million Sarobi dam project	4
USD 1.434 billion Laghman dam project	5
USD 1.094 billion Kunar and Kama dam projects	5
USAID Emergency Health and Water for Kabul project: USD 623,594	2
USAID Kabul Environmental Sanitation and Health Project: USD 4.2 million	3
USAID Kabul Water Supply and Sanitation Project: USD 20 million	3
USAID Kunar Construction of Irrigation Structure Project: USD 251,325	2
USAID Kunar Canal Cleaning and Construction of Protection Walls: USD 133,600	2
USAID Nuristan Construction of Irrigation Structures: USD 998,836	3

USAID Logar Construction of Irrigation Structures: USD 997,907	3
The river needs additional dams and flow control/management structures to improve water flow, supply, and management	-5
COMPOSITE SCORE	**69**

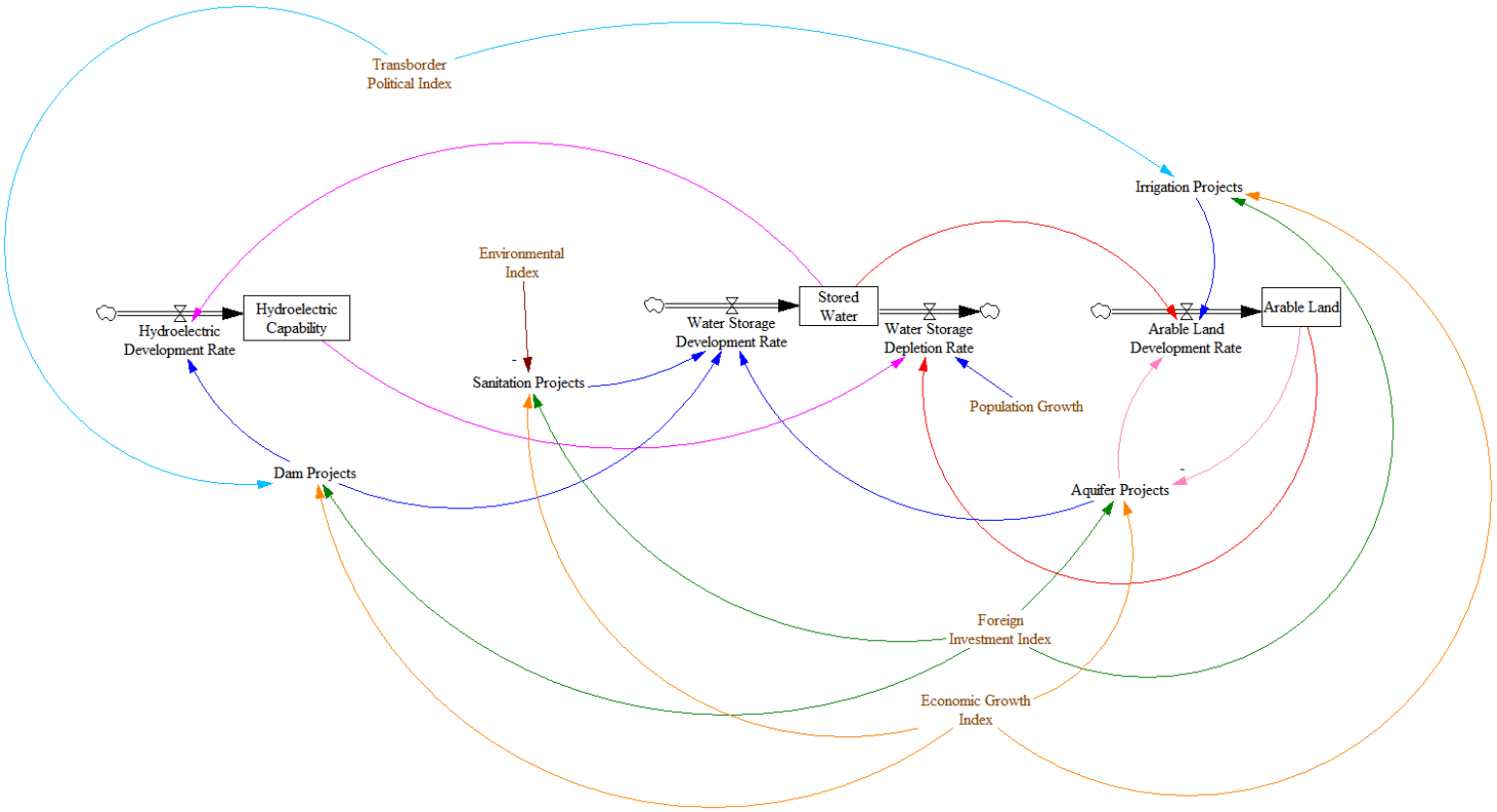
Environmental Index	
Factors	**2001-2013**
Water Stress Index: Low Risk	5
Large amount of renewable water resources (65 bcm/yr)	5
Higher control over water resources = more resilience to rainfall variability	-3
Construction of dams along rivers result in diminution of water flows and alteration of river morphology	-3
Large dam upstream effects: flooding of inhabited areas, siltation, deforestation, salinization	-3
Large dam downstream effects: severe changes to floodplain, river flow, water quality, timing and temperature, alterations to fish population	-3
Glaciers decreased by 50-70%	0
Environmental pollution reduces quality of water resources	-5
Lack of wastewater treatment procedures in populated areas results in the river carrying large amounts of diluted and floating pollutants	-5
The water quality of the Kabul River is well below international drinking standards	-3
The primary source of water is snow melt in the Hindu Kush Mountains	3
Afghanistan lacks sufficient dams, reservoirs and flow control structures to adequately manage and control this runoff. As a result, the country is susceptible to both severe flooding and droughts	-3
The Kabul River Basin contains significant areas of irrigated land	5
The Kabul Basin has a 72,000 km^2 catchment area	3
The Kabul Basin has a 22 billion m^3 storage capacity	3
Wells generally were constructed without grouting and subject to contamination from surface sources	-3
The quality of water in less populated areas was relatively good	2
During the next 50 years, a 10% reduction in total annual precipitation is anticipated in Afghanistan	-3
Increased surface temperatures in mountainous regions would be likely to result in reduced snowpacks and cause snowmelts to occur earlier in the year	-3
COMPOSITE SCORE	**-11**

Economic Growth Index	
Factors	**2001-2013**
Estimated growth rate of 6% in emerging countries	5
Rapid urbanization in developing and emerging countries	2
Water Security Risk Index: High Risk	-4
Lack of infrastructure to distribute and supply water for human consumption	-3
Lack of infrastructure to supply water for agriculture	-5
Agriculture is important part of GDP (50%)	3
Only 40% of agricultural land is currently irrigated	-5
Irrigation initiatives focus on capacity building amongst farmers who lack technical expertise and knowledge to use water efficiently	3
70% of people in Kabul City have access to electricity	5
Thy hydropower situation in Afghanistan is difficult to assess as years of conflict has left the power grids severely damaged	-5
Continued instability frequently hampers any attempts to develop the sector	-5
With 30 years of war the lack of main impediments to improved productivity is lack of investment capability, infrastructure and government institutions to provide support to farming communities	-3
Quality of planting materials either for annual or perennial crops is poor	-3
The nationwide Emergency Irrigation Rehabilitation Project (EIRP) started in June 2004 and will help farmers benefit from reliable and equitably distributed irrigation water that leads to increased agricultural productivity, better income, improved security and reduce the vulnerability of farmers to droughts	5
The average net returns per hectare of wheat production increased by 104% in Kabul	5
The river basin supports over 300,000 ha of intensively irrigated areas and high valued agricultural crops	3
Water demand in Kabul City and within the river basin is expected to increase in the future	-3
COMPOSITE SCORE	**-5**

Appendix B
Causal Loop Diagram



Appendix C
Stock and Flow Diagram



Medical & Healthcare Application

VMASC Track Chair: Dr. Andrea Parodi
MSVE Track Chair: Dr. Michel Audette

Calcium Homeostatis in a Local/Global Whole Cell Model of Permeabilized Ventricular Myocytes with a Langevin Description of Stochastic Calcium Release
Author(s): Xiao Wang, Seth Weinberg, Yan Hao, Eric Sobie, and Gregory Smith

Examining the Benefits of a 3-D Virtual Environment in Providing Psychoeducational Workshops for College Students
Author(s): Margaret Lubas and Gianluca De Leo

A Multivariate Model to Predict Endotracheal Intubation Success by Paramedics in the Out-of-Hospital Environment
Author(s): Leigh Diggs, Sameera Viswakula, and Gianluca De Leo

An Adaptive Physics-based Non-Rigid Registration Framework for Brain Tumor Resection
Author(s): Fotis Drakopoulos and Nikos Chrisochoides

Comparison of Deep Belief Neural Network versus Manifold Learning for Brain Tumor Progression Prediction
Author(s): Loc Tran, Deqi Zhou, Feng Li, and Jiang Li

A Paint-by-Numbers Active Contour-Based Approach to the Development of a Digital Brainstem Atlas
Author(s): Nirmal Patel and Dr.Michel Audette

Quality Meshing of 2D Images with Guarantees Derived by a Computer-Assisted Proof
Author(s): Jing Xu and Dr.Andrey Chernikov

Scalability of a Parallel Arbitrary-Dimensional Image Distance Transform
Author(s): Scott Pardue, Nikos Chrisochoides, and Dr.Andrey Chernikov

Multi-material Surface Extraction for Sparse Atlas-based Neuroanatomical Representation Intraoperative Tracking
Author(s): Tanweer Rashid, and Dr.Michel Audette

Modeling of Cranial Nerve Using 1-Simplex Mesh
Author(s): Sharmin Sultana and Dr.Michel Audette

Calcium homeostasis in a local/global whole cell model of permeabilized ventricular myocytes with a Langevin description of stochastic calcium release

Xiao Wang[a], Seth H. Weinberg[a], Yan Hao[b], Eric A. Sobie[c], Gregory D. Smith[a]

a. Department of Applied Science
The College of William & Mary
Williamsburg, VA 23187, USA
b. Department of Mathematics and Computer Science
The Hobart and William Smith Colleges
Geneva, NY, 14456, USA
c. Department of Pharmacology and Systems Therapeutics
Mount Sinai School of Medicine
New York, NY 10029, USA

ABSTRACT

Intracellular calcium (Ca^{2+}) signaling involves a complex interplay between global, cell-wide changes in [Ca^{2+}] and local, subcellular Ca^{2+} release events. Local signals are frequently caused by release of Ca^{2+} from intracellular stores, primarily the endoplasmic/sarcoplasmic reticulum (ER/SR). Spatially localized Ca^{2+} release events are mediated by clusters of release channels, IP_3 receptors (IP_3Rs) or ryanodine receptors (RyRs), located on the ER/SR membrane and are observable experimentally as Ca^{2+} sparks or puffs (for review see (1)).

Models of excitation-contraction (EC) coupling are able to reproduce graded Ca^{2+} release mechanistically by simulating the stochastic gating of channels in Ca^{2+} release sites using Monte Carlo methods. In these approaches, one or more L-type Ca^{2+} channels interact with a cluster of RyRs through changes in [Ca^{2+}] in a small "dyadic subspace" between the sarcolemmal and SR membranes. These models also generally consider local changes in junctional SR [Ca^{2+}], because these changes are thought to be important for Ca^{2+} spark termination and refractoriness (2–4). Realistic cellular SR Ca^{2+} release can be simulated

by computing the stochastic triggering of sparks from hundreds to thousands of such Ca^{2+} release units (CaRUs) (3–6). However, Monte Carlo simulations of local control of EC coupling can be computationally demanding, especially when each CaRU is composed of interacting Markov chain models representing the stochastic gating of individual ion channels (for review see (7)).

Here we present an alternative local/global whole cell modeling approach based on a Langevin formulation of the stochastic dynamics of Ca^{2+} release sites composed of 50–200 ryanodine receptors (RyRs). In the Langevin model, we assume that the RyRs are coupled by the change in local $[Ca^{2+}]$, and that the number of RyRs in each CaRUs is large enough that the fraction of channels in different states can be treated as a continuous variable. We show that the Langevin description of the collective gating of RyRs is a good approximation to the corresponding discrete-state continuous-time Markov chain model when the number of RyRs per release site is in the physiological range. Analytical and numerical solution of the Fokker-Planck equation (FPE) associated with the Langevin formulation validate our implementation of both approaches. Importantly, this FPE may be viewed as the master equation for a large number of identical CaRUs. By coupling the numerical solution of this FPE to balance equations for the bulk myoplasmic and network SR $[Ca^{2+}]$, a new class of local/global whole cell model is produced whose computationally efficiency scales with the number of states in the Markov chain model for an individual RyR, as opposed to the far greater number of states in a compositionally defined CaRU. The computational efficiency of the local/global whole cell model facilitates a study of Ca^{2+} homeostasis in permeabilized ventricular myocytes. We illustrate this though comparison of the minimal model to recent experiments that probe the bidirectional coupling of stochastic SR Ca^{2+} release and SR load (8). Simulations show that the effect of myoplasmic $[Ca^{2+}]$ regulating SR Ca^{2+} release is mainly mediated by regulation of RyRs. Increasing myoplasmic $[Ca^{2+}]$ results in an increase in both spark- and non-spark-mediated Ca^{2+} release. However, myoplasmic $[Ca^{2+}]$ regulates these two pathways in different ways: the spark-mediated release increases exponentially while non-spark-medicated release increase linearly as myoplasmic $[Ca^{2+}]$ increase. For a given SR $[Ca^{2+}]$, we show that two distinct steady-states can exist-one corresponding to low myoplasmic $[Ca^{2+}]$ and low release flux from junctional SR to dyadic subspace and one corresponding to high myoplasmic $[Ca^{2+}]$ and high release flux.

REFERENCES

1. Michael J Berridge. Calcium microdomains: organization and function. Cell Calcium, 40(5-6):405–12, Oct 2006.

2. M D Stern, L S Song, H Cheng, J S Sham, H T Yang, K R Boheler, and E Rios. Local control models of cardiac excitation-contraction coupling. a possible role for allosteric interactions between ryanodine receptors. J Gen Physiol, 113(3):469–89, 1999.

3. J J Rice, M S Jafri, and R L Winslow. Modeling gain and gradedness of Ca^{2+} release in the functional unit of the cardiac diadic space. Biophys J, 77(4):1871–84, 1999.

4. E A Sobie, K W Dilly, J dos Santos Cruz, W J Lederer, and M S Jafri. Termination of cardiac Ca^{2+} sparks: an investigative mathematical model of calcium-induced calcium release. Biophys J, 83(1):59–78, 2002.

5. J L Greenstein and R L Winslow. An integrative model of the cardiac ventricular myocyte incorporating local control of Ca^{2+} release. Biophys J, 83(6):2918–45, Dec 2002.

6. J L Greenstein, R Hinch, and R L Winslow. Mechanisms of excitation-contraction coupling in an integrative model of the cardiac ventricular myocyte. Biophys J, 90(1):77–91, 2006.

7. R L Winslow, A Tanskanen, M Chen, and J L Greenstein. Multiscale modeling of calcium signaling in the cardiac dyad. Ann N Y Acad Sci, 1080:362–375, Oct 2006.

8. E Bovo, SR Mazurek, LA Blatter, and AV Zima. Regulation of sarcoplasmic reticulum Ca^{2+} leak by cytosolic Ca^{2+} in rabbit ventricular myocytes. The Journal of Physiology, pages 6039–6050, 2011.

Examining the Benefits of a 3-D Virtual Environment in Providing Psychoeducational Workshops for College Students

Margaret M. Lubas[1], MSW and Gianluca De Leo[1,2], PhD
College of Health Sciences, Old Dominion University[1]
Virginia Modeling Analysis and Simulation Center[2]
mluba002@odu.edu
gdeleo@odu.edu

Keywords: 3-D virtual environment, psychoeducational workshops, reducing stigma, college students.

Abstract

College often signifies a time period when young adults begin to create their own long-term health/lifestyle behavior patterns. In spite of the high rate of stressors experienced by this population, college students are known to be hesitant to seek out support for mental health problems, and stigma is thought to be a significant deterrent from seeking out care. This article proposes the use of a 3-D virtual environment to provide psychoeducational workshops to help college students cope with stressors. The goal of the virtual environment is to utilize an interactive online setting to provide anonymous care to protect against stigma.

INTRODUCTION

College is often an adjustment period for individuals, as it marks the first time that young adults are acting autonomously in an environment that can create many sources of stress in a student's life. College students have poor sleep quality [Lund et al., 2010] high rates of anxiety and depression [Eisenberg et al, 2007; Field et al., 2012], and difficulties with relationships [Darling et al., 2007]. The stressors experienced by this demographic impact both the student's physical and mental health [Pedersen, 2012], and their overall quality of life [Dinzeo et al., 2014]. Stress management is a significant public health issue among this population, and developing strong coping mechanisms, and positive interactions with mental health services can serve as a useful lifelong relationship for students.

Despite the high rate and significant impact of stressors experienced by this demographic, the utilization rates of mental health services for college students do not match the level of reported problems [Joyce et al., 2009], as students have access to services, but often do not seek them out. Attitudes and beliefs are an important part of service utilization among college students [Downs and Eisenberg, 2012; Eisenberg et al., 2007], and stigma is often cited as a reason for not seeking treatment [Eisenberg et al., 2009].

The proposed intervention involves a multi-user virtual environment (through the Unity platform) that would allow a user to connect in a 3-D virtual world from the convenience of their own computer, smartphone, or video game system, in the location of their choice (whether it be home, work, or outside). Building a virtual meeting place (similar to a typical room used for a clinical support group) would allow for students to be connected to each other, and connected to a mental health professional, with the goal of providing comprehensive and easy access to psycho-educational workshops and support groups. The goal of this type of intervention would be to reduce stigma, as students would participate in these workshops in real time, but through the presence of avatars, making the intervention entirely anonymous. This paper outlines the design process for creating the virtual environment, and a brief overview of the future evaluative methods to assess this environment.

METHOD

The method for the proposed project involves two aspects, (1) the design of the virtual environment and the (2) experimental design to test the virtual environment. Considerations on both aspects are addressed below.

Designing the Virtual Environment

Outlining the necessary and desired functions of the virtual environment is an important first step in the design process. The goal of the virtual environment in this instance is to mimic the environment of a support group meeting place, which often consists of a room with chairs set up in a circle. A bright and welcoming environment will be used, but one that is also simple with not many environmental distractions.

Avatars are also an important aspect of the environment, and consideration must be given to their desired functions and capabilities. In this instance, because the avatars are primarily being used to simulate the presence of a group it was decided that they would not need to have any advanced movement capabilities (such as walking, moving limbs, etc.). Although manipulating facial expressions might be useful for group experiences, in this instance the avatar's primary role is to simulate a group experience, and it was decided that avatars would already be seated in the circle of chairs when participants logged in. The number of avatars in the circle would correspond to the number of participants logged into the group. Avatars could be customized based on some demographic features to provide some type of demographic information to the other group participants and facilitators, and users would be able

to move the avatars head (simulating yes or no responses) to allow for some non-verbal communication, which might further assist the group facilitator.

After the visual aspects of the environment are addressed, the functions of the environment must also be clearly outlined. In this instance, voice capabilities are an important aspect of the environment, as it would allow support group members and the facilitator to easily communicate through speech (instead of text based chat, which might be difficult for users and the group facilitator to follow). An audio track notification which would provide the participants and group facilitator with the important information of who is speaking and the help the group facilitator manage group discussions is also essential. It is also important to give the group facilitator specialized capabilities, such as the ability to mute or permanently log someone out of the group during the session (if the user becomes disruptive). Login capabilities that allows for users to login using a pc or mobile phone into a password protected group may also be beneficial. A final proposed capability of the environment is the use of a functional whiteboard/screen that displays text. This would allow the group facilitator to type in information and while it is being relayed verbally, and to have key points visually appear on a screen (similar to how this takes place in a classroom setting).

Experimental Design

After the creation of the environment, the goal of the proposed research is to test the online version of the same intervention used in a face-to-face setting to assess outcomes, student preference, and sense of presence (in the virtual meeting place). Outcomes for the comparison of delivery methods will be measured through a pre and post design to assess the achievement of learning objectives in both environments. Students will also be able to complete a post-test satisfaction survey with the workshop for both environments. Finally, sense of presence will be measured by the Igroup Presence Questionnaire, a 14 question measure that evaluates three aspects of sense of presence: spatial presence, involvement, and experienced realism [Schubert et al., 2001].

CONLUSION

Examining the effectiveness of a virtual environment in offering support to college students aims to increase service utilization among this difficult to reach population and it attempts to address their long-term psychological. If effective, virtual environments may be able to provide an infrastructure for online-based support groups and education for the college community. The virtual meeting place could be used for support groups or one-time workshops, and may serve a small group of students or a larger portion of the campus community.

In addition to its value with students, online virtual environments could also prove to be an efficient resource to military families, connecting them to specialized support across geographic areas, or individuals who live in rural areas and do not have access to services, or physically disabled individuals who are unable to leave their home. Future research should continue to explore the use of 3-D virtual environments in providing support services to individuals by examining clinical outcomes and participant satisfaction feedback. In addition to this, research should explore how to engage professionals in offering online services as past research as demonstrated professional hesitancy toward providing online interventions [Lubas and De Leo, 2014].

Reference List

Darling, C.,A. McWey, L.M., Howard, S.N., and Olmstead, S.B. 2007. "College Student Stress: the Influence of Interpersonal Relationships on Sense of Coherence." *Stress Health* 23, no.4, (October): 215-229.

Dinzeo, T., Thayasivam, U., and Sledjeski, E. 2014. "The Development of the Lifestyle and Habits Questionnaire-Brief Version: Relationship to Quality of Life and Stress in College Students." *Prevention Science Journal*, 15, no.1, (February): 103-114.

Downs, M. and Eisenberg, D. 2012. "Help Seeking and Treatment use Among Suicidal College Students." *Journal of American College Health*, 60, no.2, (February): 104-114.

Eisenberg, D., Downs, M., Golberstein, E., and Zivin, K. 2009. "Stigma and Help Seeking for Mental Health Among College Students." *Medical Care Research and Review*, 66, no.5, (October)*:* 522-541.

Eisenberg D., Golberstein E., and Gollust, S. 2007. "Help-Seeking and Access to Mental Health Care in a University Student Population. *Medical Care* 45, 7, (July): 594–601.

Field, T., Diego, M., Pelaez, M., Deeds, O., and Delgado, J. 2012. "Depression and Related Problems in University Students." *College Student Journal*, 46, *no.*1 (March): 193-203.

Joyce, A.W., Ross, M.J., Vander Wal, J.S., and Austin, C.C. 2009. "College Students' Preferences for Psychotherapy Across Depression, Anxiety, Relationship, and Academic Problems. *Journal of College student Psychotherapy*, 23, no.3: 212-226.

Lubas, M. and De Leo, G. 2014. "Online grief support groups: Facilitators' Attitudes." *Death Studies,* pending publication.

Lund, H.G., Reider, B.D., Whiting, A.B., and Prichard, J.R. 2010. "Sleep Patterns and Predictors of Disturbed Sleep in a Large Population of College Students. *Journal of Adolescent Health*, 46, no.2, (August): 124-132.

Pedersen, D.E. 2012. "Stress carry-over and college student health outcomes." *College Student Journal,*
46, no.3, (September): 620-627.

Schubert, T., Friedmann, F. and Regengrecht, H. 2001. "The experience of presence: Factor Analytic Insights." *Presence,* 10, no.3, (June): 266-281.

A Multivariate Model to Predict Endotracheal Intubation Success by Paramedics in the Out-of-Hospital Environment

Leigh Ann Diggs, MPH[a,c], Sameera D. Viswakula, MS[b], and Gianluca De Leo, PhD[a,c]
College of Health Sciences, Old Dominion University[a]
College of Sciences, Department of Mathematics and Statistics, Old Dominion University[b]
Virginia Modeling Analysis and Simulation Center[c]
ldigg004@odu.edu
sviswaku@odu.edu
gdeleo@odu.edu

Keywords: Emergency Medical Services, paramedics, advanced airway management, endotracheal intubation, patient safety

ABSTRACT

Paramedics perform life-saving procedures in the out-of-hospital setting in chaotic environments. One of the most important and controversial procedures performed by paramedics is endotracheal intubation. This extended abstract describes a study to create a multivariate model to predict successful endotracheal intubation.

1. INTRODUCTION

Paramedics provide life-saving emergency medical procedures to patients suffering from conditions such as cardiac arrest, respiratory failure, and major trauma in the out-of-hospital setting. Paramedics often have to perform advanced airway management procedures such as endotracheal intubation (ETI), the insertion of a breathing tube into the trachea (windpipe) of critically ill patients [Wang et al. 2006a]. Failure to establish a definitive airway is a major cause of preventable death when adequate oxygen and ventilation cannot otherwise be obtained [Gruen et al. 2006]. Early airway management is vital in cardiac arrest and comatose patients, as well as, patients at risk for progressive loss of airway patency and those at risk for aspiration, such as those with head and neck injuries. Oral ETI is considered the "gold standard" for definitive airway management, is regarded as one of the most important emergency medical services (EMS) procedures, and has been used in the United States for more than 25 years [Wang et al. 2006]. However, paramedic success rates in the out-of-hospital setting range from 33-100 percent [Wang and Yealy 2006, Bulger et al. 2002, Wang et al. 2003, Bulger et al. 2007]. Success rate variability has been attributed to provider differences in the level of initial training, continuing education, medical oversight, and access to neuromuscular blocking agents [Bulger et al. 2002, Bulger et al. 2007]. Due to low ETI success rates, the use of alternate airways has been recommended. A recent study characterized out-of-hospital airway management practices in the United States using the largest aggregate of EMS data collected to date, National Emergency Medical Services Information System (NEMSIS), and found a national ETI success rate of 85.3 percent. The study also found that alternate airways, except for the King (89.7%), had substantially lower success rates than ETI (Combitube (79.0%), Esophageal Obturator Airway (38%), and Laryngeal Mask Airway (66%)) [Diggs et al. 2014]. Only a few studies have been conducted using multivariate models to predict difficult ETI. A limited number of univariate studies have been conducted to predict successful ETI. The purpose of this study is to create a multivariate model to predict successful intubation.

2. METHODS

2.1 Study Design

The Institutional Review Board at Old Dominion University approved this research as an exempt study. This study will utilize 2012 EMS data from the Virginia Department of Health. This data set contains EMS data for all patients receiving endotracheal intubation in the state of Virginia for the one year period from January 1, 2012 - December 31, 2012.

2.2 Outcomes

The frequency, success rates, and complications of airway interventions will be the primary outcomes of this study. Airway interventions included in the data set include ETI (orotracheal and nasotracheal), rapid sequence intubation, and alternate airways (King LT and Combitube). The secondary outcome of this study will be a model to predict successful ETI.

2.2.1 Demographics

Characteristics, including age group, gender, race, and ethnicity, of patients receiving airway management interventions will be delineated. Cardiac arrest and possible injury will also be illustrated.

2.2.2 Complications

Complications for procedures were also contained in the data set. We will focus on complications related to airway management including anatomical abnormality, apnea, bleeding, esophageal intubation immediately detected, esophageal intubation – other, respiratory distress, vomiting, and vomitus/blood/secretions in airway.

2.2.3 Population Setting

Population setting, including metropolitan, suburban, and rural, will be analyzed to see if there are any differences in success rates by population setting.

2.3 Primary and Secondary Data Analysis

Descriptive statistics will be used to analyze the data. Univariate analysis will be conducted to see if variables, including sex, race, ethnicity, age, level of provider performing procedure, population setting, primary symptom, primary impression, myocardial infarction, first rhythm, and return of spontaneous circulation, predict ETI success. Significant ($p<0.05$) variables will then be placed in a logistic regression model to look for significance and interaction in predicting ETI success.

Detecting independent variable contributions in logistic regression begins with the following equation:

$$\text{Probability of outcome}(\hat{Y}_i) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots\beta_iX_i}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots\beta_iX_i}}$$

The linear regression equation for independent variables is expressed on a logit scale. The reasons for this logit scale transformation lies in the basic parameters of the logistic regression model. The binary outcome is expressed as a probability and must fall between 0 and 1. Logistic regression identifies the strongest linear combination of independent variables that increases the likelihood of detecting the observed outcome. This process is known as maximum likelihood estmation. To ensure an accurate model we will perform a univariate analysis of the variables and outcome first and use a direct approach to the model [Stolzfus 2011].

3. RESULTS

The results from this study will characterize airway management practices in Virginia. A multivariate model predicting ETI success will be developed, so that it can be tested using NEMSIS data and other state data.

4. DISCUSSION

The prehospital setting is an uncontrolled environment which offers unique environmental challenges that make adverse events all the more likely to occur. EMS personnel often work in small, poorly, dimly lit spaces in environments that are unfriendly, chaotic and challenging for emergent health care interventions. Unlike a hospital, emergency scenes are loud, cluttered, and unfamiliar places to out-of-hospital care providers. In addition to the environmental challenges, emotional stressors are often heightened by the presence of family members who are panicked, curious bystanders, and a lack of human and medical resources. Physical and emotional stressors are further complicated by the time sensitive nature of EMS care. The EMS arena is rich with opportunities for adverse events [Canadian Patient Safety Institute 2011].

A model of ETI success could help us determine which factors are most important in determining success. Logistic regression examines the multitude of variables under study and reveals the unique contribution of each variable after adjusting for others. This study could help us determine if provider level is, in fact, associated with ETI success. This could mean that education interventions, using techniques such as simulation, could better educate paramedics in ETI leading to a greater success rate.

REFERENCES

Bulger, E.M., Copass, M.K., Maier, R.V., Larsen, J., Knowles, J., and Jurkovich G.J. 2002. "An Analysis of Advanced Prehospital Airway Management", *Journal of Emergency Medicine*, 23: 183-189.

Bulger, E.M., Nathens, A.B., Rivara, F.P., MacKenzie, E., Sabath, D.R., and Jurkovich G.J. 2007. "National Variability in Out-of-Hospital Treatment After Traumatic Injury", *Annals of Emergency Medicine*, 49: 293-301.

Canadian Patient Safety Institute. 2009. "Patient Safety in Emergency Medical Services: Advancing and Aligning the Culture of Patient Safety in EMS", http://www.patientsafetyinstitute.ca/English/research/commissionedResearch/patientSafetyinEMS/Documents/Patient%20Safety%20in%20EMS%20Full%20Report.pdf.

Diggs, L.A., Yusuf, J.E. (Wie), De Leo, G. 2014. "An Update on Out-of-Hospital Airway Management Practices in the United States", *Resuscitation*, Article Accepted for Publication March 2014.

Gruen, R.I., Jurkovich, G.J., McIntyre, L.K., Foy, H.M. and Maier, R.V. 2006. "Patterns of Errors Contributing to Trauma Mortality: Lessons Learned form 2,594 Deaths", *Annals of General Surgery*, 244: 371-380.

Stolzfus, J.C. 2011. "Logistic Regression: A Brief Primer", *Academic Emergency Medicine*, 18, no.10: 1100-1104.

Wang, H.E., Kupas, D.F., Paris, P.M., Bates, R.R., and Yealy, D.M. 2003, "Preliminary Experience with a Prospective, Multi-centered Evaluation of Outcomes of Out-of-hospital Endotracheal Intubation", *Resuscitation*, 58: 49-58.

Wang, H.E., Lave, J.R., Sirio, C.A., and Yealy, D.M. 2006. "Paramedic Intubation Errors: Isolated Events or Symptoms of Larger Problems", *Health Affairs*, 25, no.2: 501-509.

Wang, H.E. and Yealy, D.M. 2003. "How Many Attempts are Required to Accomplish Out-of-Hospital Endotracheal Intubation?", *Academic Emergency Medicine*, 13: 372-377.

An Adaptive Physics-Based Non-Rigid Registration Framework for Brain Tumor Resection

Fotis Drakopoulos and Nikos Chrisochoides
Department of Computer Science, Old Dominion University, Norfolk, VA
fdrakopo@cs.odu.edu and nikos@cs.odu.edu

Keywords: non-rigid registration, biomechanical model, tumor resection, ITK, Finite Element Method

Abstract

We present an Adaptive Physics-Based Non-Rigid Registration (APBNRR) framework for warping pre-operative to intra-operative brain Magnetic Resonance Images (MRI) of patients who have undergone a tumor resection. The proposed method, iteratively removes the tumor from a gradually warped segmented pre-operative image via an adaptively changing biomechanical model which is necessary for dealing with deformations like those induced by a tumor resection. We show that our scheme not only accurately captures the deformations associated with the resection but also satisfies the time constraints imposed by the neurosurgical workflow. We evaluate the APBNRR framework on clinical volume MRI data and compare it with the publicly available PBNRR method of ITK. In all the case studies, our method achieves high accuracy and close to real-time performance. Indeed, APBNRR reduces the alignment error up to 6.61 and 4.95 times compared to a rigid and the PBNRR registration, respectively, while the execution time is less than 1 minute in a Linux Dell workstation with 12 Intel Xeon 3.47GHz CPU cores and 96 GB of RAM.

1. INTRODUCTION

Non-Rigid Registration between pre-operative MRI data and the in-situ shape of the brain can compensate for brain deformation during Image-Guided Neurosurgery (IGNS). Non-Rigid Registration (NRR) is a key enabling technology which brings real-time information that the surgeon is otherwise unable to collect intra-operatively.

In [1, 13] it was demonstrated that a reasonably accurate NRR of pre-operatively acquired MRI can be achieved well, within the time constraints imposed by the neurosurgical procedure, using intra-operative data. Methods [1, 4, 7] compensate for small brain deformations (shifts) caused mainly from the cerebro spinal fluid (CSF) leakage, gravity, edema and administration of osmotic diuretics. However, the complex neurosurgical procedure of brain retraction or tumor resection, which invalidates the biomechanical model defined on the pre-operative MRI and compromises the fidelity of the IGNS, is not addressed. In this paper we focus in one of those two challenges: the tumor resection.

In [8], the retraction and the resection were simulated to update the pre-operative image to realistically reflect the brain morphology in the Operating-Room (OR). A Finite Element (FE) model was created and boundary conditions were applied to the retracted surfaces. Then, the elements that coincided with the intra-operatively resected tissue were manually deleted.

In [10], an adaptive FE multi-level grid registration method which accommodates a superficial tumor resection was developed. This method evaluated only in 2D medical and synthetic images. In [9], a robust Expectation-Maximization (EM) framework was presented to simultaneously segment and register a pair of 3D clinical images with partial or missing data. A MatLab implementation of this method required 30 min to register a pair of $64 \times 64 \times 64$ volumes on a 2.8 GHz Linux machine.

In this paper, we augment the software implementation in [7] and propose an Adaptive Physics-Based Non-Rigid Registration (APBNRR) framework to compensate for the brain deformation induced by a tumor resection. The proposed scheme removes automatically the tumor from a gradually warped segmented pre-operative image, while an adaptive biomechanical model deals with the complex brain deformations occurring during the resection. Our method is reasonably fast to satisfy the time constraints required by the neurosurgical procedure. We introduce several parallel components, thus we can register adult brain MRIs with resolution $250 \times 219 \times 176$ voxels in less than 60 seconds. The evaluation of our framework is based on 6 volume clinical cases with : (i) brain shifts (2 cases), (ii) partial tumor resections (2 cases) and (iii) complete tumor resections (2 cases). In all the case studies, the APBNRR achieves higher accuracy compared to the publicly available non-rigid registration method PBNRR [7] of ITK[1] and exhibits close to real-time performance.

In the next section we will describe the proposed scheme that manages the FE model adaptivity. A comprehensive description of the framework with an extensive evaluation on a larger data set is available at [6].

[1] http://www.itk.org/

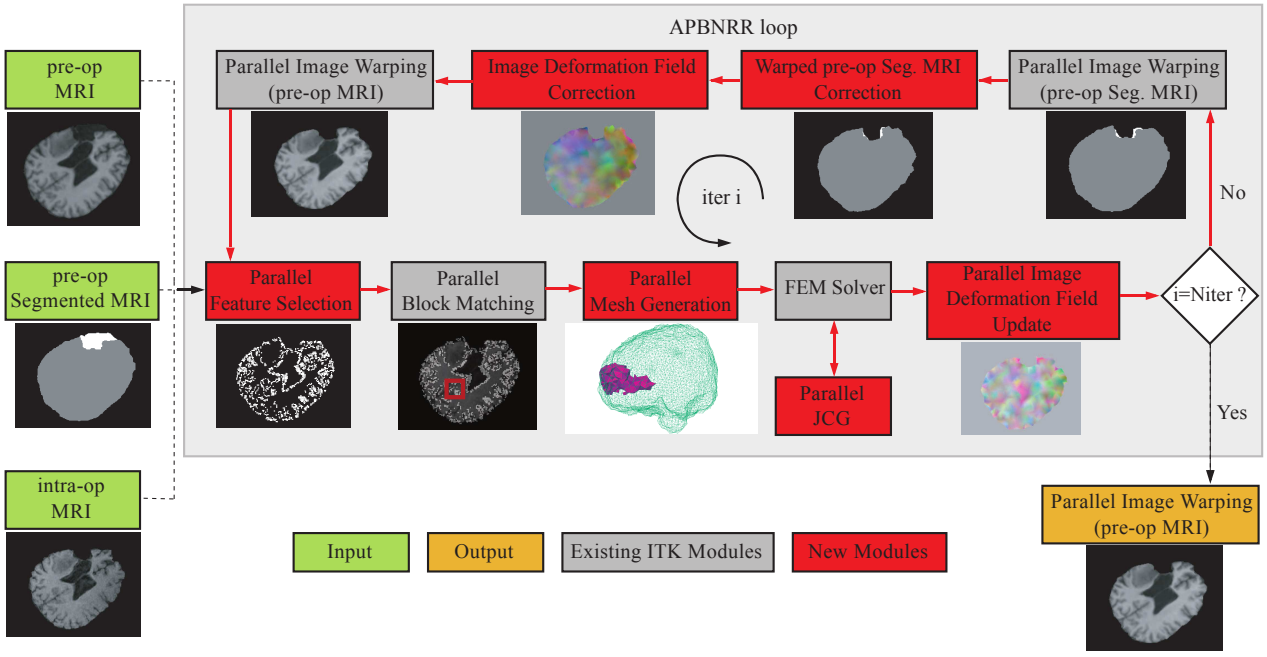


Figure 1. The APBNRR framework [6]. The green, red and gray boxes represent the input, the new contributions and the existing ITK modules, respectively. The red arrows show the execution order of the modules. Orange represents the output warped pre-operative MRI.

2. METHOD

The APBNRR framework is built on the ITK open-source system. Figure 1 illustrates the modules of the framework. All parallel modules are developed with the POSIX thread library.

The basic idea of the APBNRR method is to iteratively estimate a dense deformation field that defines a transformation for every point in the intra-operative to the pre-operative image. The estimation of the dense field is facilitated by a heterogeneous (brain parenchyma, tumor) FE biomechanical model of high quality tetrahedral elements. During the execution, the model deforms and adapts to the new brain morphology induced by the tumor resection (Figure 1).

In each APBNRR iteration, first we select high discriminant features (blocks) from the warped pre-operative MRI (when $i = 1$ the warped pre-operative MRI equals to the input pre-operative MRI). Then, we compute a sparse displacement field that matches the selected features to their corresponding blocks in the intra-operative MRI (block matching displacements). Next, we apply the sparse field of matches to the model and we estimate the deformations on the mesh vertices with the solution of a linear system of equations. The model stiffness and consequently the computed mesh deformations mostly depend on: a) the mechanical properties of the brain and tumor tissues, b) the shape (quality) of the elements, c) the number and the positions of the selected blocks, d) the block matching displacements.

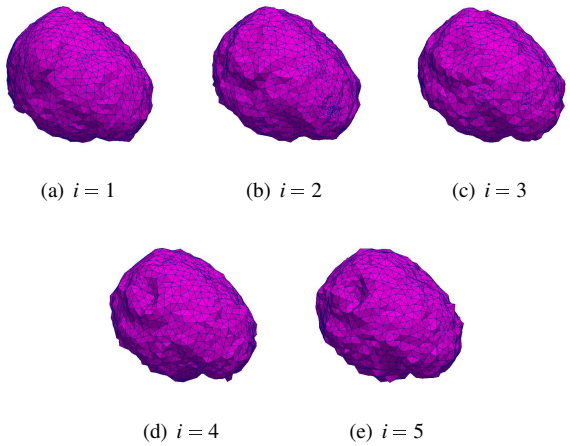
In a later step, we convert the mesh deformations to an image deformation field which is used to warp the pre-operative and the segmented pre-operative images. Additionally, we apply correction modules on the image deformation field and the warped segmented pre-operative image, to compensate for the resected tissue (Figure 1). We should point out that the image deformation field is additive; it holds the sum of the previous image fields at iterations $1, 2, \ldots, i-1$ and the current image field at iteration i. In that way, independently of the number of iterations, we interpolate only the input pre-operative and segmented pre-operative images.

Figure 2 shows the FE brain model adaptivity implemented on the APBNRR framework. The example consists of five adaptive iterations. For each mesh: a) its surface is conformed to the segmented image boundary of the current iteration i, and b) the distorted poor quality tetrahedral elements occurring after each deformation are eliminated.

The model deformation and consequently the image warping, stops when $i = Niter$, where $Niter$ is the desired number of adaptive iterations (Table 2). Our experimental evaluation has shown that a satisfactory alignment accuracy can be achieved within the neurosurgical time constraints, when $Niter = 3-5$. The output registered image is the warped pre-operative MRI at iteration $Niter$ (Figure 1). A complete description of the new and existing APBNRR modules can be found in [6, 7].

Table 2. The input parameters for the 6 clinical cases. BS : Brain Shift, PTR : Partial Tumor Resection, CTR : Complete Tumor Resection, FS : Feature Selection, BM : Block Matching, MG : Mesh Generation, FEMS : FEM Solver, All : PBNRR-APBNRR, x : axial, y : coronal, z : sagittal.

Parameter	Units	Value	Description	Module	Method
$B_{s_x} \times B_{s_y} \times B_{s_z}$	voxels	$3 \times 3 \times 3$	Block size	FS-BM	All
$W_{s_x} \times W_{s_y} \times W_{s_z}$	voxels	$7 \times 7 \times 7$ (BS)	Window search size	BM	All
		$9 \times 9 \times 9$ (PTR, CTR)			
F_s	-	5%	% of selected feature blocks	FS	All
δ	-	5	Mesh size	MG	APBNRR
E_b	Pa	2.1×10^3	Brain Young's modulus	FEMS	All
E_t	Pa	2.1×10^4	Tumor Young's modulus	FEMS	APBNRR
ν_b	-	0.45	Brain Poisson's ratio	FEMS	All
ν_t	-	0.45	Tumor Poisson's ratio	FEMS	APBNRR
λ	-	1	Trade off parameter	FEMS	All
F_r	-	25%	% of rejected outlier blocks	FEMS	All
N_{appr}	-	10	Number of approximation steps	FEMS	All
N_{int}	-	5	Number of interpolation steps	FEMS	All
N_{iter}	-	3 (BS)	Number of adaptive iterations	-	APBNRR
		4 (PTR,CTR)			



(a) $i = 1$ (b) $i = 2$ (c) $i = 3$

(d) $i = 4$ (e) $i = 5$

Figure 2. The adaptive FE biomechanical model implemented in the APBNRR framework. Each mesh is conformed to the warped segmented pre-operative image of iteration i. Number of generated tetrahedra for $i = 1 - 5$: $7725, 8102, 7720, 7262, 6991$.

3. RESULTS

We evaluate our framework on 6 clinical volume MRI cases and we compare it with the publicly available non-rigid registration method PBNRR [7] of ITK. Prior to the non-rigid registration we extract the brain from the skull with BET [11] and we rigidly align the pre-operative to the intra-operative MRI with 3D Slicer[1]. All MRI data are anonymized and an Institutional Review Board (IRB) is granted. The Surgical Planning Laboratory at Brigham and Women's Hospital [12] provided the first three cases and the Department of Neurosurgery at Shanghai Huashan Hospital provided the last three [3]. Depending on the type of resection depicted in the intra-operative MRI (i.e., just brain shift but no tumor resection, or partially/completed resected), the cases are categorized as Brain Shifts (BS), Partial Tumor Resections (PTR) and Complete Tumor Resections (CTR). From totally 6 cases, 2 are BS, 2 are PTR and 2 are CTR. Table 1 lists the provided clinical data. All MRI data were resampled to a uniform image spacing $1.00 \times 1.00 \times 1.00$ (mm) along the x, y, z (axial, coronal, sagittal) image directions. For all the conducted experiments we used linear displacement FE biomechanical models with 4-node tetrahedral elements and the tissues (brain parenchyma, tumor) were modeled as elastic isotropic materials. Table 2 lists the parameters for the experiments. More details about the parameters are given in [6, 7].

[1] http://www.slicer.org/

Table 1. The clinical MRI data of this study. BS : Brain Shift, PTR : Partial Tumor Resection, CTR : Complete Tumor Resection.

Case	Type	Provider	Genre	Tumor Location
1	BS	B&W	M	R frontal
2	BS	B&W	F	R occipital
3	PTR	B&W	F	L frontal
4	PTR	Huashan	M	L frontal
5	CTR	Huashan	M	R temporal
6	CTR	Huashan	F	L posterior temporal

3.1. Quantitative evaluation

For the quantitative evaluation, we employ the Hausdorff Distance (HD) metric as it is implemented in [5].

The HD is computed between extracted point sets in the warped pre-operative and the intra-operative images. For the point extraction we employ ITK's Canny edge detection method [2]. We compute the alignment errors HD_{RIGID}, HD_{PBNRR} and HD_{APBNRR}, after a rigid, a non-rigid (PBNRR) and an adaptive non-rigid (APBNRR) registration, respectively. The smaller the HD value, the better the alignment. Additionally, we compute the alignment improvement of the APBNRR compared to the rigid and the PBNRR registration. The corresponding ratios are HD_{RIGID}/HD_{APBNRR} and HD_{PBNRR}/HD_{APBNRR}. When ratio > 1 the APBNRR outperforms the other method. The higher the ratio, the greater the improvement.

In Table 3 we present the quantitative results. Figure 3 depicts all HD values and their corresponding average values for all the experiments. As shown in Table 3 and Figure 3, our method, in all the case studies, significantly reduces the alignment error compared to the rigid and the PBNRR registration. The maximum improvement occurs in case 5 (CTR), with values 6.61 and 4.95, respectively (Table 3). On the average the APBNRR is 4.23 and 3.18 times more accurate than the rigid and the PBNRR registration, respectively (Table 3). Generally, the APBNRR performs better on the PTR and CTR cases, because it captures accurately the large, complex intra-operative deformations associated with the tissue resection. On the other hand, the PBNRR is a non-adaptive method which is designed to handle only the small brain shifts occurring during the surgery.

Table 3. The quantitative evaluation results for the 6 clinical cases. $HD_{RIGID}, HD_{PBNRR}, HD_{APBNRR}$ is the alignment error after a rigid, a non-rigid (PBNRR) and an adaptive non-rigid registration (APBNRR), respectively. All HD are in mm.

Case	HD_{RIGID}	HD_{PBNRR}	HD_{APBNRR}	$\frac{HD_{RIGID}}{HD_{APBNRR}}$	$\frac{HD_{PBNRR}}{HD_{APBNRR}}$
1	13.63	11.22	5.91	2.30	1.89
2	8.60	7.00	5.38	1.59	1.30
3	19.33	16.03	3.74	5.16	4.28
4	12.72	9.43	2.82	4.51	3.34
5	16.15	12.08	2.44	6.61	4.95
6	19.62	12.53	3.74	5.24	3.35
Average	15.00	11.38	4.00	4.23	3.18

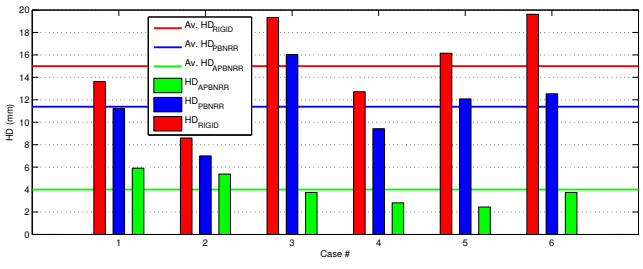


Figure 3. The Hausdorff Distance (HD) error for the 6 clinical cases. The horizontal lines illustrate the average HD error.

3.2. Qualitative evaluation

Figure 4 depicts the qualitative results for cases PTR (3-4) and CTR (5-6). These cases clearly demonstrate the impact of our method on the challenging problem of tumor resection. Figure 4 shows the same representative slice for all the MRI belonging to the same row. The cyan color delineates the tumor segmentation in the pre-operative image. The fifth and sixth column (from the left) show the warped pre-operative MRI subtracted from the intra-operative MRI. The black and white regions in the difference images indicate larger discrepancies, while the gray regions indicate smaller discrepancies. Obviously, the APBNRR aligns the images with high accuracy, particularly near the tumor resection margins where the black and white regions are mostly eliminated. Moreover, the APBNRR provides accurate alignments independently of the portion of the resected tissue depicted in the intra-operative image (partial or complete tumor resection). On the contrary, the PBNRR cannot compensate for the large deformations induced by the resection and shows significant misalignments nearby the tumor cavities.

3.3. Performance evaluation

In this paper we perform all the experiments in a Dell Linux workstation with 12 Intel Xeon X5690@3.47GHz CPU cores and 96 GB of RAM. Figure 5 shows the total (end-to-end) APBNRR execution time, for all the case studies, with 1, 4, 8, and 12 hardware cores. Because of the various implemented multi-threaded modules, our method is able to register the clinical data in less than 1 minute (between 34.51 and 56.17 seconds), as shown with green in Figure 5. We should point out that the APBNRR does not scale linearly with the number of the cores. There is a significant speed boost from 1 to 4 cores, but limited improvement from 4 to 12 cores. The reason is mainly that APBNRR has not fully parallelized yet (Figure 1), so the maximum achieved speedup is always limited by Amdahl's law. After the parallelization of the sequential modules (Figure 1) and especially the computationally intensive FEM Solver [6], we expect to reduce the end-to-end execution time by 30-40% and achieve the alignments in less than 45 seconds.

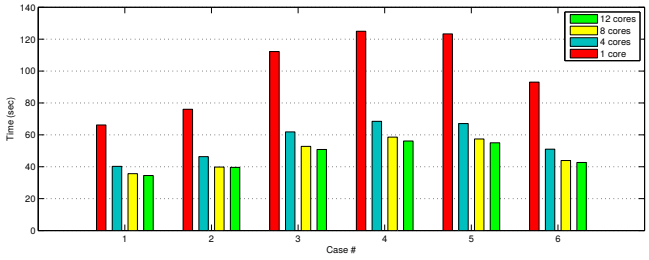


Figure 5. The APBNRR (end-to-end) execution time for the 6 clinical cases using 1, 4, 8, and 12 hardware cores.

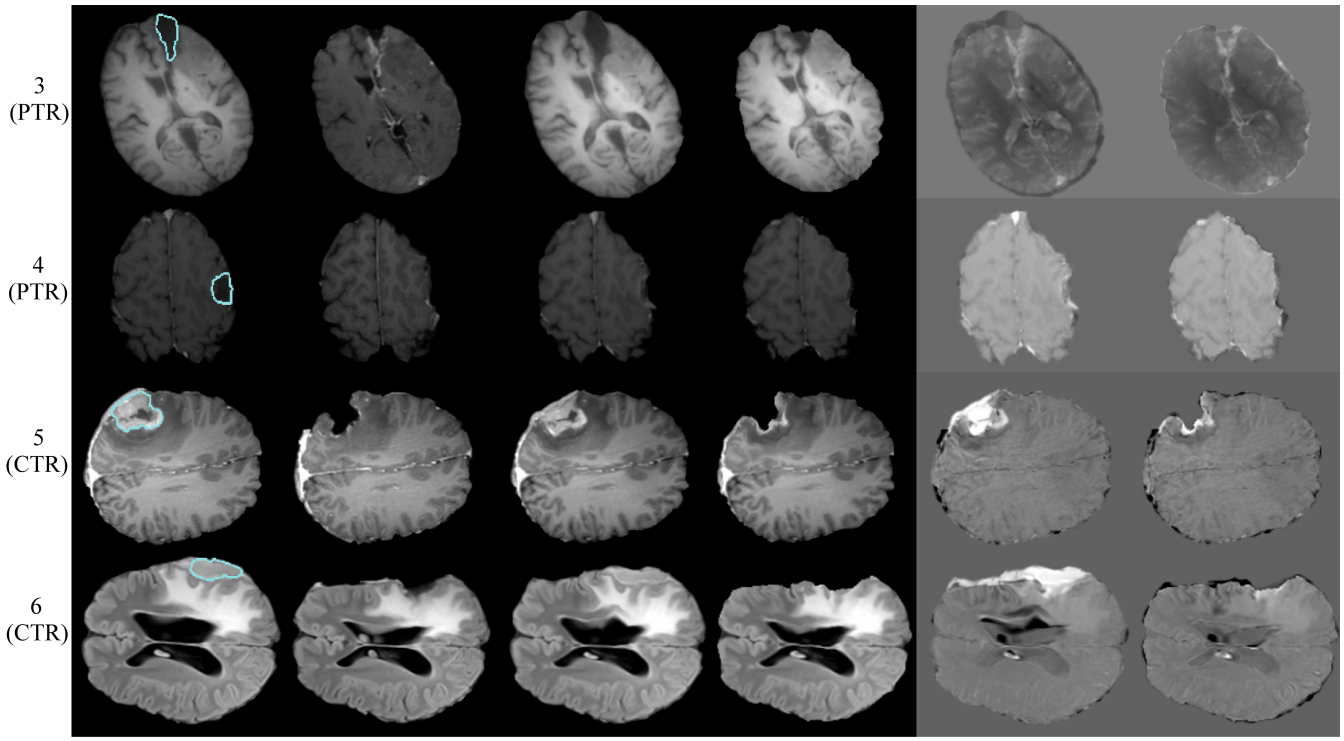


Figure 4. Qualitative evaluation results for the tumor resection cases. Each row represents a single case. The left margin indicates the number and the type of each case. From left to right column: pre-op MRI, intra-op MRI, warped pre-op MRI (PBNRR), warped pre-op MRI (APBNRR), warped pre-op MRI (PBNRR) subtracted from intra-op MRI, warped pre-op MRI (APBNRR) subtracted from intra-op MRI. For the PTR and CTR cases the cyan color delineates the tumor segmentation in the pre-op MRI.

4. SUMMARY AND CONCLUSION

We presented an Adaptive Physics-Based Non-Rigid Registration (APBNRR) framework to compensate for the brain deformations induced by a tumor resection.

The proposed method is built on the ITK open-source system and implements an adaptively changing heterogeneous (brain parenchyma, tumor), patient-specific, FE biomechanical model, to warp the pre-operative to the intra-operative MRI. We show that our framework can accurately handle the complex brain deformations associated with the neurosurgical procedure, independently of the portion (partial/complete) of the resected tissue depicted in the intra-operative MRI.

Our evaluation is based on clinical volume MRI data from 6 patients acquired from two hospitals. In all the conducted experiments our scheme exhibited high registration accuracy. It reduced the alignement error up to 6.61 and 4.95 times, compared to a rigid registration and the publicly available non-rigid registration method PBNRR of ITK, respectively.

Besides, most of the APBNRR modules are parallel. In all the case studies we tried in a Dell Linux workstation with 12 Intel Xeon X5690@3.47GHz CPU cores, our method needed between 34.51 and 56.17 seconds to register a pair of volume MRIs. Consequently, our scheme fits well within the time constraints (less than 1-2 minutes) imposed by the neurosurgery procedure. For this reason and considering the high accuracy of the provided alignments, we believe that our method has a potential use in the Operating-Room.

In the future, we will incorporate more tissues (e.g. brain ventricles) into the model in order to improve the registration accuracy. Also, we will parallelize the sequential modules to reduce further the end-to-end execution time.

REFERENCES

[1] N. Archip, O. Clatz, A. Fedorov, A. Kot, S. Whalen, D. Kacher, N. Chrisochoides, F. Jolesz, A. Golby, P. Black, and S. K. Warfield. Non-rigid alignment of preoperative mri, fmri, dt-mri, with intra-operative mri for enchanced visualization and navigation in image-guided neurosurgery. *Neuroimage*, 2007.

[2] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, January 1986.

[3] Wang Chen, Mao Ying, Zhu Jian-Hong, and Zhou Liang-Fu. The department of neurosurgery at shangai huashan hospital. *Neurosurgery*, 62(4):947–953, April 2008.

[4] O. Clatz, H. Delingette, I.-F. Talos, A. Golby, R. Kikinis, F. Jolesz, N. Ayache, and S.K. Warfield. Robust non-rigid registration to capture brain shift from intra-operative mri. *IEEE Trans. Med. Imag.*, 2005.

[5] F. Commandeur, J. Velut, and O. Acosta. A vtk algorithm for the computation of the hausdorff distance. *The VTK Journal*, 2011.

[6] Fotis Drakopoulos, Yixun Liu, Panagiotis Foteinos, and Nikos P Chrisochoides. Towards a real time multi-tissue adaptive physics based non-rigid registration framework for brain tumor resection. *Frontiers in Neuroinformatics*, 8(11), 2014.

[7] Yixun Liu, Andriy Kot, Fotis Drakopoulos, Chengjun Yao, Andrey Fedorov, Andinet Enquobahrie, Olivier Clatz, and Nikos P Chrisochoides. An itk implementation of a physics-based non-rigid registration method for brain deformation in image-guided neurosurgery. *Frontiers in Neuroinformatics*, 8(33), 2014.

[8] M.I. Miga, D.W. Roberts, F.E. Kennedy, L.A. Platenik, A. Hartov, K.E. Lunn, and K.D. Paulsen. Modeling of retraction and resection for intraoperative updating of images. *Neurosurgery*, 49(1):75–84; discussion 84–5, 2001.

[9] Senthil Periaswamy and Hany Farid. Medical image registration with partial data. *Medical Image Analysis*, 10(3), 2006.

[10] Petter Risholm, Eigil Samset, Ion-Florin Talos, and William Wells. A non-rigid registration framework that accommodates resection and retraction. In *Information Processing in Medical Imaging*, volume 21, pages 447–458, 2009.

[11] Stephen M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[12] I-F. Talos and N. Archip. Volumetric non-rigid registration for mri-guided brain tumor surgery. 08 2007.

[13] Simon K. Warfield, Matthieu Ferrant, Xavier Gallez, Arya Nabavi, and Ferenc A. Jolesz. Real-time biomechanical simulation of volumetric brain deformation for image guided neurosurgery. In *Proceedings of the 2000 ACM/IEEE conference on Supercomputing*, Supercomputing '00, Washington, DC, USA, 2000. IEEE Computer Society.

Comparison of Deep Belief Neural Network versus Manifold Learning for Brain Tumor Progression Prediction

Loc Tran[1], [2]Deqi Zhou, Feng Li[1], Jiang Li[1]

[1]ECE, Old Dominion University, Norfolk, VA 23508
[2]The IB Program at Princess Anne High School, Virginia Beach, VA 23456

Abstract. We compare two data centric approaches for brain tumor progression prediction, namely manifold learning and deep neural network. Manifold learning models identify low-dimensional structures embedded in high dimensional data sets. Computational restrictions usually limit these models on large scale data sets such as the tumor progression data set. Deep neural networks are an advanced form of artificial neural network with multiple hidden layers. The data sets consist of a series of high dimensional MRI scans for four patients with tumor and progressed regions identified. By identifying and modeling tumor progression, these methods have the potential to greatly benefit patient management.

1 Introduction

Overtime, advancements in medical imaging technology have allowed for the acquisition of multimodal, large-scale, and heterogeneous medial datasets. For example, brain magnetic resonance (MR) imaging exams now incorporate MR diffusion tensor imaging (DTI) on a frequent basis. Furthermore, this imaging modality, along with traditional T1, T2, or FLAIR weighted MRI scans give more practical information as well as the potential to provide a more qualitative brain tumor diagnosis [1]. But with the large amount of data, a challenge exists in interpreting these large-scale and high-dimensional data sets.

Within the past few years, predicting glioma tumor progression has been extensively studied using mathematical models. Microscopic, Mesoscopic, and Macroscopic are three categories these mathematical models are typically classified in. Microscopic models use sub-cellular levels of data by focusing on the tumor cell internally to describe the growth process. Mesoscopic models focuses on the relationship between tumor cells and surrounding tissue [2]. Macroscopic models note macroscopic quantities, such as tumor volume and blood flow, while focusing on tissue level processes. Furthermore, most macroscopic methods use Murray's diffusion equation for reaction-diffusion modeling [3]. These models usually consist of predicative values and a set of estimated parameters. In this paper, we investigate two data-centric methods, namely manifold learning and deep neural networks, for brain tumor progression prediction that use only information from high-dimensional MRI scans.

2 Method

2.1 Data Preparation

MRI data was collected using FLAIR, T1-weighted, post-contrast T1-weighted, and DTI MRI scans from four patients with progressing brain tumors. Five scalar volumes which include apparent diffusion coefficient (ADC), fractional anisotropy (FA), max eigenvalues, middle eigenvalues, and min eigenvalues were also computed from the DTI volume and yields a total of ten image volumes for each visit per patient. Each patient

went through a series of clinical visits over the course of two years. Using the vtkCISG toolkit, a strict form of registration was applied to each patient that aligned all volumes to the DTA volume [4]. After registration,

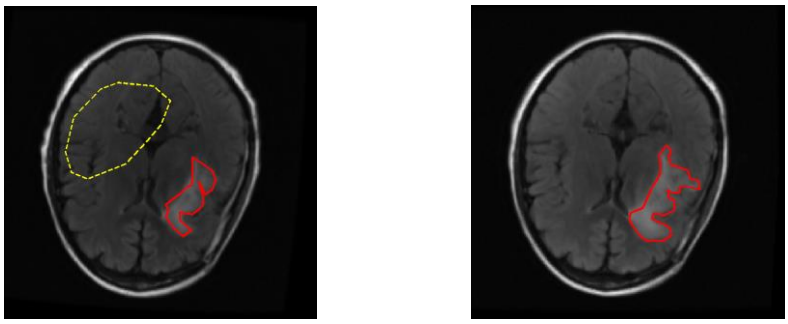


Figure 1. Tumor and normal regions defined for Subject 1 where the red tumor regions are labeled by a radiologist and the yellow polygon denotes normal regions. FLAIR images at Visit 1 (left) and Visit 2 (right) showing a progressed tumor at visit 2.

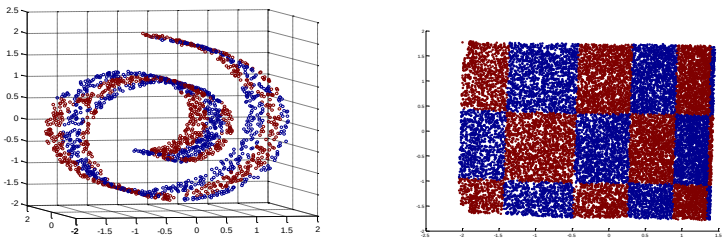


Figure 2. Example unfolding of a nonlinear manifold. (left) Structure in 3 dimensions. (right) 2-dimensional representation.

each pixel location can be represented by a ten-dimensional feature vector corresponding to the ten MRI scans. One visit was selected and labeled as "Visit 1," and a later visit that showed evidence of tumor progression was selected and labeled "Visit 2." A radiologist defined the tumor regions on the FLAIR scans. For training purposes, normal regions far from the tumor regions have also been defined. Figure 1 shows the selected regions for one subject for both visits. The yellow dotted region is the normal region while the red polygon is the abnormal region. Note that the abnormal region is larger at Visit 2. The goal of this study is to predict the progressed region at Visit 2.

2.2 Proposed methods

Manifold learning approaches attempt to find low dimensional and non-linear structures embedding within high dimensional data. These methods then unfold the manifolds and represent them in a low dimensional linear subspace. While these methods are effective in deciphering nonlinear structures, the low-dimensional representation is difficult to compute, especially for large data sets. Figure 2 shows a graphical example of manifold learning. In this case, the plot on the left shows the data in three dimensions. The intrinsic structure of the data is only in two dimensions. If the geometry is unrolled, the dataset can be represented as the right plot which uses only the intrinsic dimensionality. Extracting the intrinsic geometry is computationally intensive, mainly because of an n x n spectral decomposition of a similarity matrix where n is the size of the data set. For large data sets such as the MRI data in this study, a direct computation of the eigen-decomposition is avoided through sampling [5-7]. This allows for the manifold to be calculated using a smaller number of points. But since all of the points are not used to calculate the manifold, the result can be seen as

only an approximation of the manifold learned on the entire data set. Deep belief neural networks are an extension on the standard feed forward neural network. The standard model consists of an input layer and an

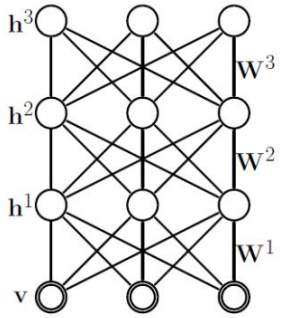


Figure 3. Structure of deep belief neural network. Bottom layer is the input. Upper layers are hidden layers, *h*, with weights, *W*, between them.

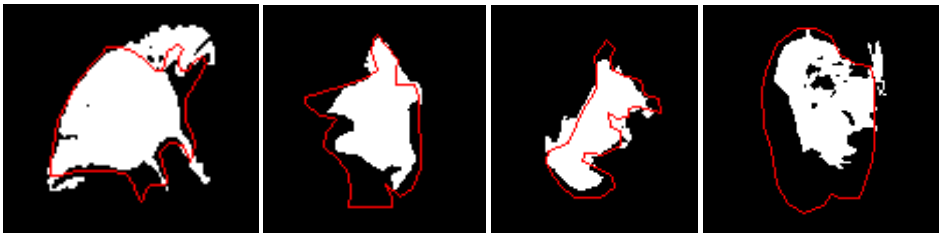


Figure 4. Cropped images of abnormal classification region for four subjects using deep neural network. The red outline denotes radiologist marked regions for Visit 2.

output layer with one hidden layer between whereas a deep neural network has multiple hidden layers [8]. An example structure of a deep neural network is given in Figure 3. The first layer is still the input layer. In this experiment, we implemented three hidden layers with 100, 25, and 5 hidden units respectively. The last layer consists of 2 outputs corresponding to a normal and abnormal classification. Each layer is connected to the previous layer with a set of weights. The challenge of all neural networks is to calculate the these weights. The weights connecting each layer are pre-trained using a restricted Boltzmann machine (RBM). Fine-tuning of weights is performed using back propagation.

3 Experiments and Results

Both models were trained on a sampled set from both the abnormal and normal region at Visit 1. The output of the neural network will be a classification prediction. For manifold learning, a Gaussian mixture model (GMM) classifier is applied to the low dimensional manifold. Testing points are embedded into the manifold using local linear embedding. Classification of testing points are then dictated by the GMM classifier.

Table 1 show quantitative performance metrics calculated as an average over 4 subjects. The sensitivity measures the ratio between the number of pixels correctly predicted as abnormal versus the total number of marked abnormal pixels. This measure was calculated for both Visit 1 and Visit 2. Specificity is the ratio of the correctly predicted normal tissue samples inside the normal contours. The precision is the number of correctly predicted abnormal pixels divided by the total number of predicted abnormal points. The precision will be 1 if

every pixel predicted as abnormal is within the marked abnormal region and conversely, the metric will be low for methods that have an over-estimated tumor region. The precision was calculated only at Visit 2 because the abnormal region was expected to expand between Visit 1 and Visit 2. The results for RAW are found by

Table 1. Average sensitivities and specificities for the four patients

	Sensitivity at Visit 1	Specificity at Visit 1	Sensitivity at Visit 2	Precision at Visit 2	Average
Deep Neural Network	0.948	1.000	0.653	0.858	**0.864**
Manifold Learning	0.951	1.000	0.663	0.781	0.849
PCA	0.945	1.000	0.649	0.473	0.766
RAW	0.917	0.872	0.705	0.617	0.778

directly applying a GMM classifier in the high dimensional space. For PCA, the dimensionality reduction is performed using principal component analysis.

By comparing each method, it can be seen that the deep neural network method is an overall better method than the other techniques for this data set. The sensitivity and specificity are comparable with the manifold learning approach while the precision is considerably higher. This means that the predicted abnormal points are more likely to correspond to the actual progression region. Figure 4 shows the classification region for deep neural network for each of the four patients. The red outline is the marked abnormal regions at Visit 2.

4 Conclusion

We show that a more robust tumor model can be achieved using an deep belief neural network compared to large scale manifold learning. Both the manifold learning and deep neural network produced better results compared to using raw data and PCA. This suggests that the tumor growth model is nonlinear in nature.

References

[1]. Bode, M. K., J. Ruohonen, et al. "Potential of Diffusion Imaging in Brain Tumors: A Review." Acta Radiol, no. 47: 585-594, 2006.

[2]. Hatzikirou, H., A. Deutsch, et al., "Mathematical Modelling of Glioblastoma Tumor Developement: A Review." *Mathematical Models and Methods in Applied Sciences*, 15(11): 1779-1794, 2005.

[3]. Murray, J., Mathematical Biology. Heidelberg, Springer, 1989.

[4]. T. Hartkens, D. Rueckert, J. A. Schnabel, D. J. Hawkes, and D. L. G. Hill, "Vtk cisg registration toolkit: An open source software package for affine and nonrigid registration of single- and multimodal 3d images.," in *Bildverarbeitung fur die Medizin* (M. Meiler, D. Saupe, F. Kruggel, H. Handels, and T. M. Lehmann, eds.), vol. 56 of CEUR Workshop Proceedings, pp. 409–412, Springer, 2002.

[5]. A. Talwalkar, S. Kumar, and H. Rowley, "Large-scale manifold learning," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8, 2008.

[6]. L. Tran, D. Banerjee, X. Sun, J. Wang, A. J. Kumar, D. Vinning, F. D. McKenzie, Y. Li, and J. Li, "A large-scale manifold learning approach for brain tumor progression prediction," in *Proceedings of the Second international conference on Machine learning in medical imaging*, MLMI'11, (Berlin, Heidelberg), pp. 265–272, Springer-Verlag, 2011.

[7]. Deshpande, A., L. Rademacher, et al., "Matrix approximation and projective clustering via Volume Sampling." *In Symposium on Discrete Algorithms*,no. 2, pp. 225-247, 2006.

[8]. G. E. Hinton and R. R. Salakhutdinov, " Reducing the Dimensionality of Data with Neural Networks" Science, 313(5786): 504-507, 2006. [DOI:10.1126/science.1127647]

A paint-by-numbers active contour-based approach to the development of a digital brainstem atlas

Nirmal Patel, Michel A. Audette

Abstract

Increasingly, there is a requirement for MRI segmentation in terms of functional regions, where the complexity of this functional map entails a digital atlas approach, rather than a voxel or surface-based approach, which cannot disambiguate a large number of tissues due to overlapping intensities. This requirement holds true for patient-specific anatomical modeling for neurosurgical planning and simulation, where for example various neuroanatomical targets are indicated in deep-brain stimulation. There is an insufficiency of descriptive digital atlases, while a large number of printed atlases could be leveraged to address this need. One specific example of an unmet need for a digital functional map, relevant to our work, is a digital brainstem atlas. This presentation will focus on an on-going work on a level-sets (LS) approach to transposing 2D images of a printed atlas of the brainstem, starting from a numbered, contour-based sketch of a slice of the brainstem and proceeding to fill in each corresponding label, which is analogous to painting numbered regions in a paint-by-numbers kit.

We adopt a piecewise 2D labeling approach, where each label in the scanned atlas page is used to initialize a rectangular contour, which is propagated outwardly to coincide with the contour of the region. In general, a gradient-based LS model is used. In order to preprocess the image, anisotropic diffusion filter is applied to smoothen the image to minimize the irregularities within regions which can affect the contour evolution. In addition, the atlas may contain arrows which may cross regions. These arrows are removed by initializing two points at each end of each arrow. The minimal path, which stays inside the arrow, between the two points is then extracted. However, since the minimal path is thin, a LS model is used to evolve the extracted path as a starting contour to cover the entire arrow. As regions are often colored, a selective median filter which ignores the neighboring pixels inside the arrow is applied on each arrow pixel. This removes the arrow and replaces it with the color of the nearby pixels. Some atlases may also have open regions. In order to keep the contour from diffusing outside the open region, lines may need to be drawn manually. This is done by inputting two end points for each line.

Each functional map elaborated thus, defined in the x-y plane, will then be inserted at its coordinate on the z-axis and registered with neighboring images. This analysis will produce a volumetric brainstem model suitable for mapping to patient MRI data, which will then facilitate cranial nerve modeling.

Quality Meshing of 2D Images with Guarantees Derived by a Computer-Assisted Proof

Jing Xu and Andrey N. Chernikov
Department of Computer Science,
Old Dominion University,
Norfolk, VA, USA,
{jxu,achernik}@cs.odu.edu

Keywords: mesh generation, angle bounds, computer-assisted proof, interval arithmetic

Abstract

This paper describes an algorithm for generating unstructured triangular meshes from a continuous two-dimensional object represented by an image. The algorithm uses squares as a background grid from which to build the quadrilateral elements that conform to the input contours. Then the triangulation is obtained by automatically choosing the diagonals that optimize the angles of the triangles. The extracted triangular meshes can be extensively used in the finite element method (FEM) since our triangulation provides a minimum angle bound of 18.4349°. The angle bound is verified by a computer-assisted proof using interval arithmetic.

1. INTRODUCTION

An important challenge of engineering decision-making is to establish procedures that can represent the physical reality with sufficient accuracy to make predictions. The main procedures are indicated schematically in Fig. 1.

Our work mainly concerns with the procedure of discretization, in other words, mesh generation. There are usually two kinds of meshes used in finite element methods and its applications. The first is surface mesh, which explicitly represents the surface of an object [2]. The second kind of mesh, called volumetric mesh, is distinct from surface mesh in that it explicitly represent both the surface and the volume of the structure [11]. We are developing a three-dimensional tetrahedral mesh generation algorithm, this paper is a preliminary result. We describe our two-dimensional algorithm with full details in this paper.

Finite element methods are now an important and frequently indispensable part of engineering analysis and simulation and modeling. Finite element computer programs are now widely used in practically all branches of engineering for the analysis of fluids, interfaces, and solids. The first step in the finite element computation is to discretize the problem domain into a union of elements, this step is often termed mesh generation. A common choice for an element in two dimensions is the triangle, in three dimensions is the tetrahedron.

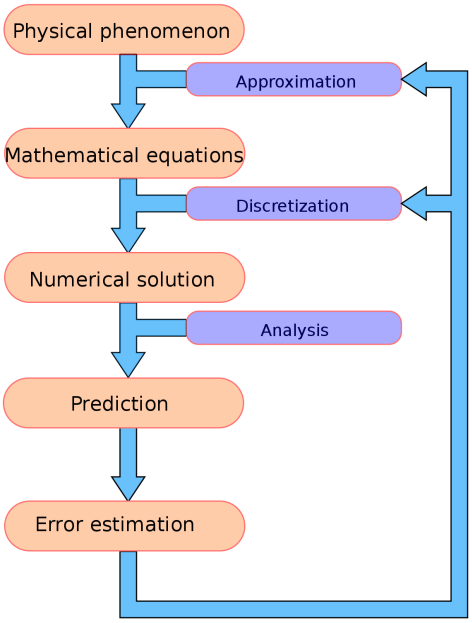


Figure 1. The main procedures of numerical simulation and error estimation.

Thus, a triangulation of the domain is required.

The quality of meshes as well as the number of mesh elements influences the accuracy of the finite element method and condition number of the stiffness matrix. Just a few bad elements can ruin the whole computation. On triangular meshes, the discretization errors usually occur when large angles approach 180°; and both large and small angles are deleterious to the stiffness matrix conditioning [10]. The size of mesh elements influences the computation time and memory storage; the memory requirement and computation time for the numerical process increases drastically when the number of elements increases.

There has been a significant amount of algorithms that theoretically guarantee a quality mesh. In principle, Delaunay triangulation, advancing front, and finite quadtree method (octree for three-dimensional) are all applicable to two or three-dimensional mesh generation.

Mesh generation by Delaunay refinement, whose input is

planar linear complexes, is one of the push-button algorithms used for constructing guaranteed quality triangular and tetrahedral meshes. Quality is traditionally defined in terms of the bounds on circumradius-to-shortest-edge ratio [3]. The use of this measure leads to the improvement of the minimum angle in two dimensions, which helps to improve the conditioning of the stiffness matrix used by a field solver. In three dimensions this measure does not yield such direct benefits.

The advancing front algorithm, taking the input models defined by level set function, generates elements one by one from an initial 'front' formed from the specified boundary of the domain, until the whole domain is completely covered by elements [7,9]. Usually this method accompanies with mesh quality enhancement techniques, such as mesh smoothing and mesh modification.

Quadtree method, whose input is also a level set function, was introduced for domain decomposition to generate non-uniform meshes by Yerry and Shephard [12]. Later, Bern, Eppstein, and Gilbert [1] studied several versions of generating triangular meshes of a planar point set or polygonally bounded domain which guarantee well-shaped elements and small total size simultaneously (Mitchell and Vavasis [8] extended Bern's work to three dimensions). Labelle and Shewchuk [6] adopted the idea of warping and proposed the Isosurface Stuffing tetrahedral meshing algorithm on geometric domains represented by a continuous cut function.

This paper develops a fast triangular mesh generation method. The algorithm generates a triangulation for smooth bounded domain with or without holes. In addition, our method offers two guarantees. First, all the elements in the meshes generated by our algorithm have high quality, meaning that all the angles of all the triangles are bounded between 18.4349° and 143.1302°. Second, the number of triangles is within a constant factor of the best possible for any triangulation with bounded angles. Besides, it is numerically robust and simple to implement. It is applicable in any numerical simulation of the partial differential equations such as fluid flow, mechanical deformation, and diffusion.

The angle bounds were obtained through a computer-assisted proof. These bounds hold for any continuous cut function without sharp edges or corners. Interval arithmetic is used in the proof to get the conservative bounds.

A second version of our algorithm creates meshes whose interior triangles are graded, but on the boundary, the triangles have uniform size. The algorithm relies on a balanced quadtree subdivision that offers interior grading. Since it ensures that the mesh elements are uniform on the objects' boundary, the angle bounds for the uniform meshes also apply to the graded ones.

The remainder of this paper is organized as follows. In Section 2 we give more details on our uniform meshing algorithm. In Section 3 we describe the boundary graded quality mesh generation algorithm. In Section 4 we provide the optimality proof. Section 5 concludes the paper.

2. UNIFORM MESHING ALGORITHM

Our algorithm starts with constructing a initial regular mesh from which to construct an output mesh. Then our algorithm identifies all the edges and points that across the boundaries of objects. During the third step, we deform the initial regular mesh so that a set of initial regular mesh edges respect objects' boundary. And finally we obtain the output mesh by chosing the best triangulation.

2.1. Physical Domain and Background Mesh

We first define a bounded domain $\Omega \subset R^2$, where R is the set of reals. Then we define $f : R^2 \to R$ be a continuous level set function that implicitly represents the geometric shape of the physical domain. The points in point set $\{p : f(p) = 0\}$ are on the boundaries of the domain. Points where f is negative are inside the domain; points where f is positive are outside the domain, and usually should not be meshed. Fig. 2 shows an example of the level set function, an ellipse.

The algorithm employs a space-tiling initial regular mesh to guide the creation of the output mesh. Let $L := Z^2$ be a square lattice, i.e., the set of points whose coordinates are integers. Let s be the lattice spacing, we denote the uniformly scaled square lattice by $L(s)$, where all two dimensions are scaled by $s > 0$.

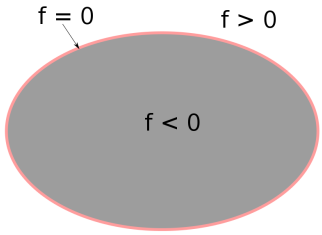


Figure 2. An example of level set function, an ellipse.

2.2. Identify the Cut Edges and Cut Points

For each lattice point of the initial regular mesh, our algorithm computes the sign of the value of the function. For any point inside the level set, the sign is assumed negative, and is assumed positive if the point lies outside the level set. If a lattice point just happen to lie exactly on the level set, the value is zero.

We define a *cutedge* E to be an edge in the initial regular mesh such that it has different signs at two end points. If an edge is a *cutedge*, there exists one point that exactly lies on both the edge and the level set. We define this kind of points a

cut point c such that $c \in E$, and $f(c) = 0$. Then the algorithm identifies all the edges that cross the level set and for each one of those edges a cut point is computed. In Fig. 3, cut edges are shown in red and cut points are shown in blue.

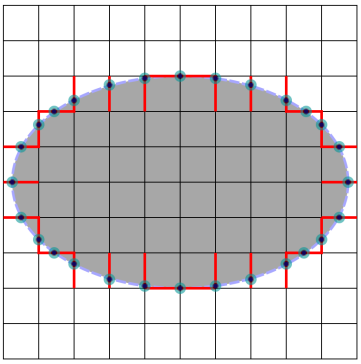


Figure 3. Cut edges and cut points.

2.3. Warping

The idea of deforming the initial regular mesh to conform the objects' boundary is to select a subset of mesh vertices that approximate the level set and to force these mesh vertices snap to the boundary. These vertices are chosen to prevent an interior mesh vertex from being connected to an exterior vertex through a mesh edge. The destination where these vertices be warped is the location of cut points. When one end point is warped to the boundary, the sign of the cut function no longer be positive or negative, it becomes exactly zero.

Let v be an end point of a cut edge E, let $D \in R$ be the distance portion function between v and the cut point c lying on this edge, and let l be the edge length function, then D can be calculated as:

$$D(v,c,E) = \frac{|v-c|}{l(E)} \tag{1}$$

We always choose the one of the two end points to warp whose Euclidean distance is shorter. That means,

$$D(v,c,E) \leq 0.5 \tag{2}$$

If one mesh vertex can be warped to several cut points, we choose any one of them. Fig. 4 shows the initial regular mesh after warping.

2.4. Triangulating the Background Mesh by Optimization

In general, after endpoints of all cut edges are warped to the destination cut points, there will be quadrilaterals and squares remaining in the initial regular mesh. We triangulate these quadrilaterals and squares by selecting different diagonals such that our choice optimizes the minimum angle in the two triangles we obtain. Fig. 5 illustrates the rule for choosing the diagonal.

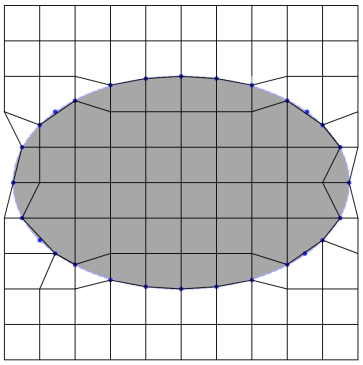


Figure 4. Initial lattice after warping.

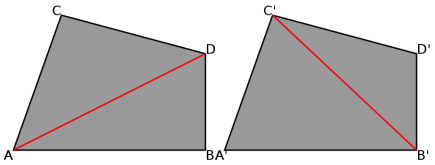


Figure 5. Rule of choice of the diagonal. The quadrilateral $ABCD$ and the quadrilateral $A'B'C'D'$ are the same quadrilateral. The minimum angle in quadrilateral $ABCD$ is $\angle BAD$, and the minimum angle in quadrilateral $A'B'C'D'$ is $\angle B'C'D'$. Because $\angle BAD > \angle B'C'D'$, we chose diagonal AD instead of diagonal $B'C'$. If the minimum angles in two quadrilaterals are same, choose anyone of them.

After choosing the diagonal, a set of triangles are obtained from the initial regular mesh. But only the ones inside the level set should be part of the final mesh. The following condition checks which triangle should be selected.

Let v_0, v_1 and v_2 be the three vertices of a candidate output triangle, let *iso* be the isovalue, the triangle is output ed if and only if

$$(f(v_0) \leq iso) \wedge (f(v_1) \leq iso) \wedge (f(v_2) \leq iso) \tag{3}$$

Fig. 6 shows the final mesh of the ellipse.

2.5. Pseudocode

Fig. 7 summarizes the process:

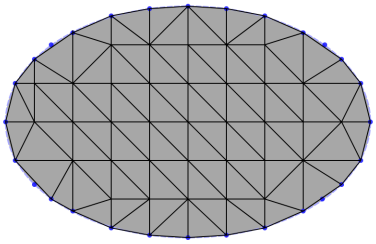


Figure 6. Final mesh.

UNIFORM MESH GENERATION(f, Iso)
Input: f is the level set function
iso is the isovalue specified by user
Output: triangular mesh $\mathcal{M}$ that represents the geometry feature of an object
1: Construct initial regular mesh B which is composed of a set of uniform squares. The size of squares is defined by user
2: Find all cut edges and their associated cut points
3: Warp, i.e., move endpoints of all cut edges to the destination cut points
4: Triangulate the warped initial regular mesh by selecting the best diagonals which optimize the minimum angle in each initial regular mesh elements
5: Output mesh

Figure 7. Pseudocode of uniform Mesh Generation.

3. GRADED MESHING ALGORITHM

Another version of our algorithm creates meshes that have graded elements in the interior but uniform fine elements on the boundary. The reason for this is for many applications, the need of accuracy on the boundary is greatest and most crucial, but the need in the interior does not have the same importance. Thus, the computational time for finite element method can be reduced by reducing the number of elements in the meshes. Fig. 8 illustrates an example of a graded interior mesh.

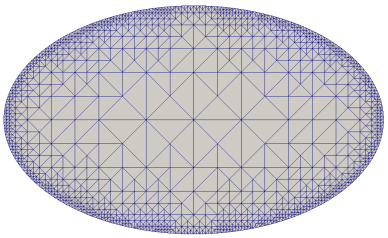


Figure 8. A graded interior mesh.

3.1. Pseudocode

We first present the pseudocode for our graded interior meshing algorithm in Fig. 9:

GRADED INTERIOR MESH GENERATION(f, Iso)
Input: f is the level set function
iso is the isovalue specified by user
Output: Graded mesh $\mathcal{M}$ that represents the geometry feature $\mathcal{M}$ has the fine uniform elements on its boundary
1: Construct initial regular mesh B by quadtree subdivision, the size of minimum leaves is defined by user
2: Balance it by requiring that any two neighboring squares differ at most by a factor of two in size
3: Find all cut edges and their associated cut points
4: Warp, i.e., move endpoints of all cut edges to the destination cut points
5: Triangulate the warped initial regular mesh by two rules. First, if at least one of the leaf side is split by a midpoint, introduce the center of the leaf and connect the center to the midpoint and the endpoints of the shared side; Second, select the best diagonals which optimize the minimum angle in each initial regular mesh elements
6: Output mesh

Figure 9. Pseudocode of graded interior mesh generation.

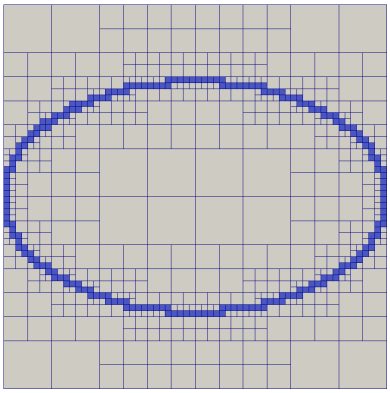


Figure 10. A balanced quadtree with marked leaves of an ellipse.

3.2. Quadtree Subdivision

The main difference between the uniform meshing algorithm and the graded interior meshing algorithm is the first step, i.e., constructing the initial regular mesh, although the implementation for the latter is more complicated. The main data structure we use for the graded meshing algorithm is a quadtree. The quadtree subdivision starts at subdividing a square, we call it a box. We commonly refer to each node of the quadtree as a sub-box. Later sub-boxes are warped and triangulated, changing their geometric structure. An quadtree node is either a leaf, or has four children. The process of generating the four children of a node is called *splitting*. A quadtree subdivision is a recursive function which needs a condition to terminate.

To make sure that small angles never be created in the mesh, after the quadtree subdivision, we balance it by requiring that any two neighboring squares differ at most by a fac-

tor of two in size. Fig. 10 shows the balanced quadtree with marked leaves that cross ellipse's boundary.

3.3. Optimizing the Triangulation

We triangulate quadtree leaves using the following two strategies: first, we say that the side of a sub-box is *split* if either of the neighboring sub-boxes sharing it is split. If at least one of its sides is split by a midpoint, introduce the center of the sub-box and connect the center to the midpoint and the endpoints of the shared side. Second, we select the mesh edge by optimizing the minimum angle in the triangulation as Fig. 11 illustrates. If none of the four sides was split, we select the best diagonal following the same rule in the uniform algorithm.

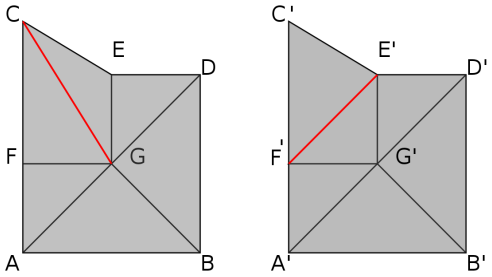


Figure 11. An example of the use of the rule for chosing mesh edges. The north side and the west side are splinted by midpoints. The sub-box $ABCD$ and the sub-box $A'B'C'D'$ are the same box. Vertex C and vertex C' are warped vertices. The triangulation of sub-box $ABCD$ is set $\{\triangle ABG, \triangle BDG, \triangle AFG, \triangle CFG, \triangle CEG, \triangle DEG\}$; and the triangulation of sub-box $A'B'C'D'$ is set $\{\triangle A'B'G', \triangle B'D'G', \triangle A'F'G', \triangle C'F'E', \triangle F'E'G', \triangle D'E'G'\}$. We compare the minimum angle in $\triangle CFG$, $\triangle CEG$ and the minimum angle in$\triangle C'E'F'$, $\triangle E'F'G'$, if the former is larger than the later, we choose edge CG as the remaining mesh edge; or choose edge $E'F'$ otherwise.

4. EXPERIMENTAL RESULTS

We applied the algorithm to a variety of shapes and images. All the steps in the previous two sections were implemented in C++. The input data is a scalar function $f(x,y)$ (i.e., a level set function). All tests were performed on a desktop with two Intel Core Xeon CPU with 3.06 GHz and 64 GB of main memory.

4.1. Geometric Shapes

Fig. 13 shows a breakdown of the total run time of graded interior algorithm applied to a sphere model into the major computational parts, as the diameter of the sphere grows from 100 to 400 pixels. These parts are the computation of a balanced quadtree, warping, and triangulating the initial regular mesh. For the simple shapes, $f(x,y)$ is implemented analytically. Fig. 12 shows a breakdown of the total run time of uniform algorithm applied to a sphere model into the major computational parts, as the diameter of the sphere grows from 100 to 400 pixels. These parts are the computation of the initial regular mesh, warping, and triangulating the initial regular mesh. We exclude the time taken by input and output.

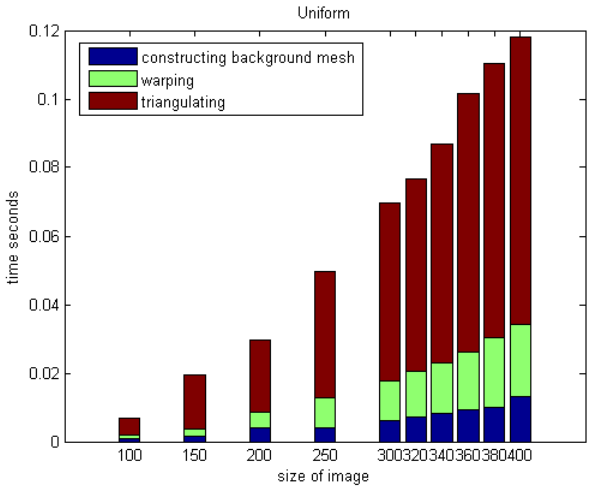


Figure 12. A breakdown of the total run time of uniform algorithm into the major computational parts, as the diameter of the sphere varies from 100 to 400 voxels.

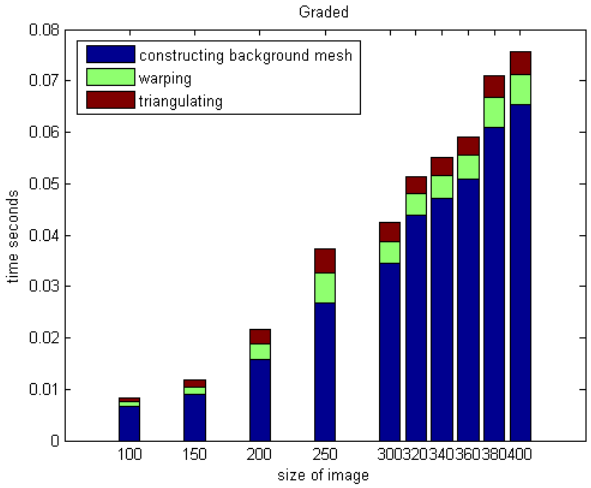


Figure 13. A breakdown of the total time of graded interior algorithm into the major computational parts, as the diameter of the sphere varies from 100 to 400 voxels.

4.2. Slices of Three-dimensional Images

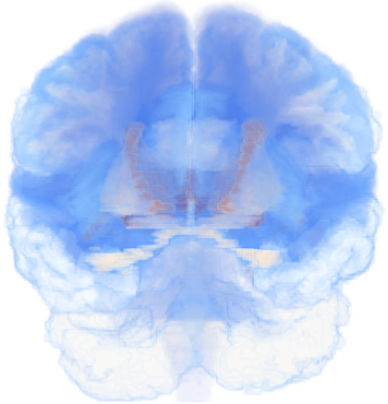


Figure 14. Original three-dimensional image of the brain atlas.

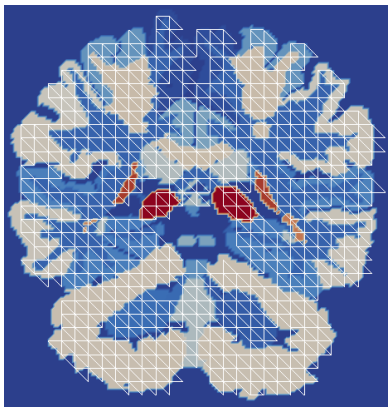


Figure 15. Uniform coarse mesh generated from background lattice with 2500 squares.

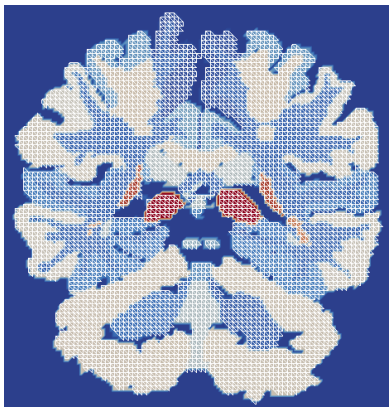


Figure 16. Uniform fine mesh generated from background lattice with 22500 squares.

We used two complex real-world medical images: slices from an abdominal atlas [5] and slices from a brain atlas [4]. The atlases come with a segmentation, such that each voxel is assigned a unique label. These slices are sampled by 256×256 rectilinear grids. The level set function $f(x,y)$ was defined by linear interpolation. The original images and output meshes are shown in Fig. 14 to Fig. 20. The running time and output mesh sizes are given in Table 1.

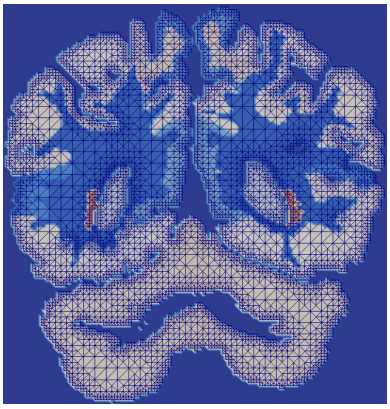


Figure 17. Graded mesh with fine boundary generated from background lattice with balanced quadtree of one slice of brain atlas.

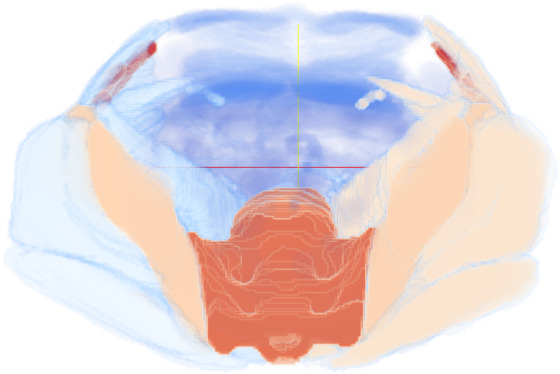


Figure 18. Original three-dimensional image of the abdominal atlas.

Table 1. Run time and size in output meshes

Output Mesh	Run Time	Number of Triangles
Mesh in Fig. 15	2.310391(s)	843
Mesh in Fig. 16	20.82288(s)	8336
Mesh in Fig. 17	46.22675(s)	14489
Mesh in Fig. 19	12.07399(s)	3132
Mesh in Fig. 20	44.50765(s)	12056

5. GUARANTEES FOR THE OUTPUT MESH

Our algorithm offers a guarantee on the output meshes that it never generates triangles with bad angles. Specifically, the

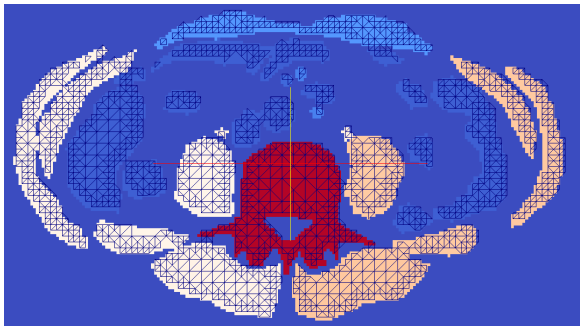


Figure 19. Graded mesh with coarse boundary generated from background lattice with balanced quadtree of one slice of abdominal atlas.

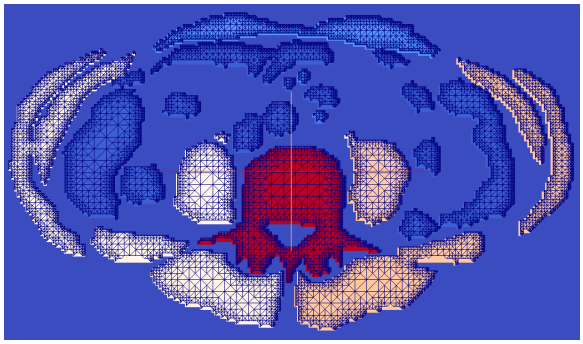


Figure 20. Graded mesh with fine boundary generated from background lattice with balanced quadtree of one slice of abdominal atlas.

angles in output mesh are bounded between 18.4349° and 143.1302°.

The main idea of our meshing algorithm is to ensure that the mesh elements on the objects' boundary are uniformly generated, so the angle bounds apply to our uniform meshes as well as graded ones. Our minimum angle bound was obtained through a computer-assisted proof. By placing a lower bound on the smallest angle of a triangulation, we are also bounding the largest angle, since in two dimensions, if no angle is smaller than θ , then no angle is larger than $180 - 2\theta$.

In our proof we work with a single square which represents any quadrilaterals in the initial regular mesh. We call it a generic square because our proofs are valid for any combination of locations the corners could be warped to. Since the number of locations where a cut point might be placed is infinite, we use interval arithmetic to verify the angle bounds. We divide the intervals of possible triangle configurations into a finite number of subintervals in that interval arithmetic calculates conservative bounds.

As illustrated in Fig. 21, each corner could lie on four segments, along x direction and y direction. We break each of the segment into n intervals, thus each corner could be located in $4n$ intervals. So the square requires the analysis of 4^{4n} cases.

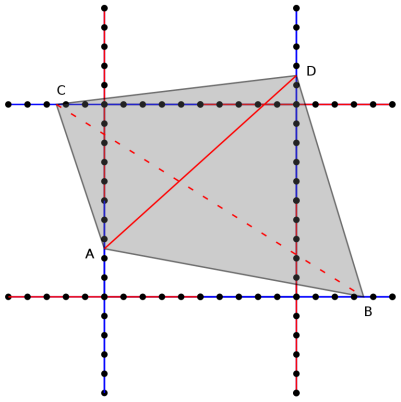


Figure 21. Gray quadrilateral is the generic square warped to position A, B, C and D. Because the minimum angle in $\triangle ABC$ and $\triangle BCD$ is smaller than the minimum angle in $\triangle ACD$ and $\triangle ABD$, we choose diagonal AD instead of diagonal BC.

In each of those cases, we choose the best diagonal optimizing the minimum angle in this square. Our minimum angle bound is obtained among all the minimum angles in those cases.

6. CONCLUSION

In this paper, we presented our new meshing algorithm both for uniform and graded mesh generation. We provide the angle bounds that make the resulting meshes suitable for FE simulation. Moreover, we proved that our mesh elements is optimal, and the meshes we created respect the geometric shapes of the input objects. For the future work, we are planning to extend it to three-dimensional tetrahedral mesh generation and its quality proof.

REFERENCES

[1] M. Bern and J. Gilbert. provably good mesh generation. In *Proceedings of the 31st Annual IEEE Symposium on the Foundations of Computer Science*, pages 231–241, New York, 1990.

[2] Andrey Chernikov and Jing Xu. A computer-assisted proof of correctness of a marching cubes algorithm. In *International Meshing Roundtable*, pages 505–523, Orlando, FL, October 2013. Springer.

[3] L. P. Chew. Guaranteed-Quality Triangular Meshes. Tech. Rep. pages TR–89–983, Department of Computer Science, Cornell University, 1989.

[4] I. Talos M. Jakab R. Kikinis and M. Shenton. Spl-pnl brain atlas, 2008.

[5] I. Talos M. Jakab R. Kikinis and M. Shenton. Spl abdominal atlas, 2010.

[6] F. Labelle and J. R. Shewchuk. Isosurface stuffing: Fast tetrahedral meshes with good dihedral angles. In *ACM Transactions on Graphics, special issue on Proceedings of SIGGRAPH 2007*, volume 26(3), pages 57.1–57.10, August 2007.

[7] S.H. Lo. *A new mesh generation scheme for arbitrary planar domains.* Int. J. Numer. Meth.Eng., 1985.

[8] S. A. Mitchell and S. A. Vavasis. Quality mesh generation in three dimensions. In *the ACM Computational Geometry Conference*, pages 212–221, 1992.

[9] J Peraire and M Vahdati. *Adaptive remeshing for compressible flow computations.* J. Comp. Phys., 2002.

[10] Jonathan Richard Shewchuk. *What Is a Good Linear Finite Element? Interpolation, Conditioning, Anisotropy, and Quality Measures.* Preprint, 2002.

[11] Jing Xu and Andrey Chernikov. A guaranteed quality boundary graded triangular meshing algorithm backed by a computer-assisted proof. In *International Meshing Roundtable*, Orlando, FL, October 2013. Springer. 5-page research note.

[12] Mark A. Yerry and Mark S. Shephard. A modified quadtree approach to finite element mesh generation. *Computer Graphics and Applications*, 3:39–46, 1983.

Scalability of a Parallel Arbitrary-Dimensional Image Distance Transform

Scott K. Pardue, Nikos P. Chrisochoides, Andrey N. Chernikov
Old Dominion University Computer Science Department
spardue@cs.odu.edu, nikos@cs.odu.edu, achernik@cs.odu.edu

Keywords: Parallel computing, parallel distance transform, distance transform, scalability, speedup, efficiency, algorithms, Euclidean Distance Transform, EDT, medical imaging, image processing

Abstract

Computing the Euclidean Distance Transform (EDT) for binary images is an important problem with applications involving medical image processing, computer vision, computational geometry, and pattern recognition. Currently, there exists a sequential algorithm of O(n) complexity developed by Maurer et al. and a parallel implementation of Maurer's algorithm developed by Staubs et al. with a theoretical complexity of O(n/p) for n voxels and p threads. In this paper, we present an efficient, scalable parallel implementation of Maurer's algorithm for large datasets with high efficiency for 16 cores.

1. Introduction

An EDT of an N-dimensional binary image is an N-dimensional matrix representing the Euclidean distance of each volume pixel (voxel) in the binary image to the closest foreground element in the binary image. Currently, the best sequential algorithm for calculating the EDT of a binary image, developed by Maurer[1], computes it in linear time. Maurer's algorithm focuses on dimensionality reduction by computing the EDT at each dimension by generating partial Voronoi diagrams for each row of voxels in each dimension. This is done by first initializing the EDT by iterating through each row of voxels in the binary image for the lowest dimension. When a foreground voxel is found, the associated voxel in the EDT is set to zero while all others are set to infinity. The voxel data of the binary image is only used for initialization. The EDT is used for the remainder of the computations. After initialization, each row of voxels for each dimension in the EDT is iterated through beginning with the lowest dimension. All voxels that are equal to infinity are disregarded. All of the remaining voxels, known as feature voxels (FV) are compared against the two closest FV to determine if the current FV intersects the row of voxels that is currently being examined. If the current FV does not intersect the current row, then this FV is disregarded. After all of the FV have been compared, the row of voxels is iterated through comparing the current Euclidean distance for the given voxel to the Euclidean distance to each FV. The lower of the two distances is the new Euclidean distance for the current voxel. The EDT of the lower dimension is used to calculate the EDT of the current dimension. The order of processing for a two dimensional image is shown in Figure 1. Each row, of the first dimension is iterated through, then each row of the second dimension is iterated through.

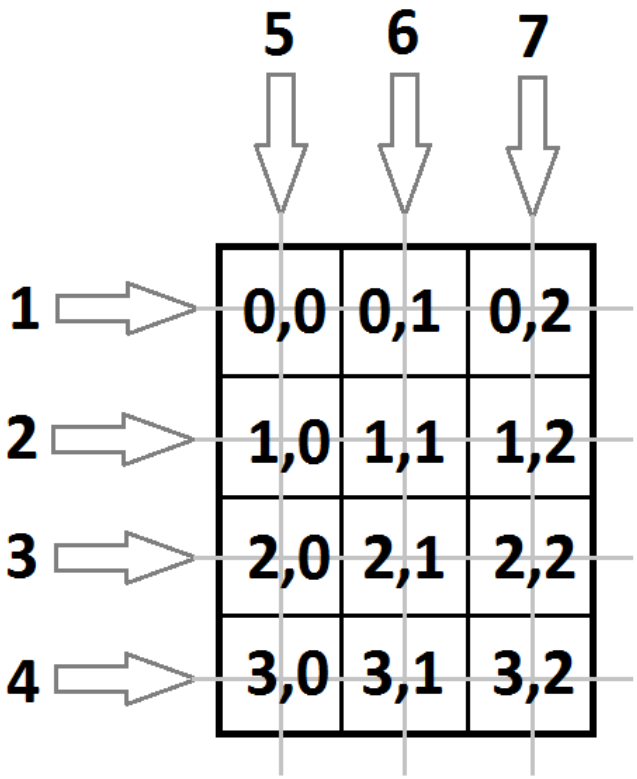


Figure 1. Order of Rows to Process

Figure 2 and Figure 3 show an example of an image of brain ventricles from the Brain Atlas[3] and its corresponding 3D distance transform, respectively.

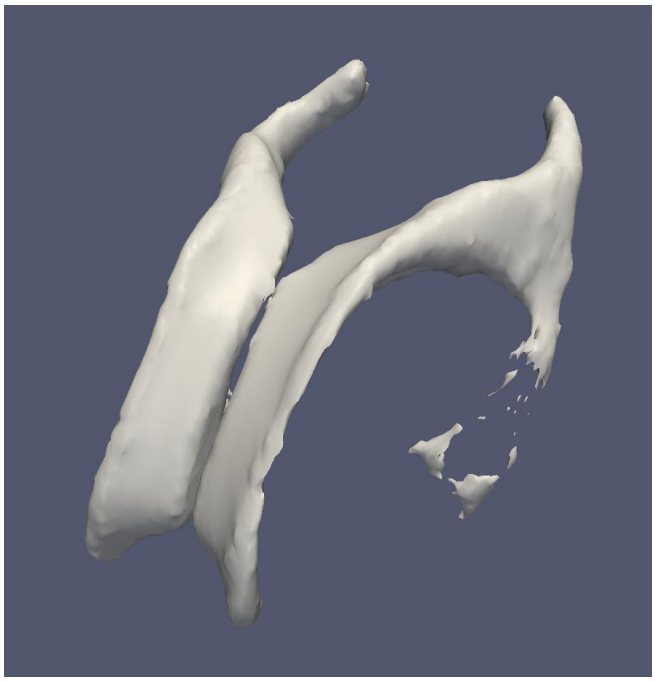


Figure 2. Image of Brain Ventricles

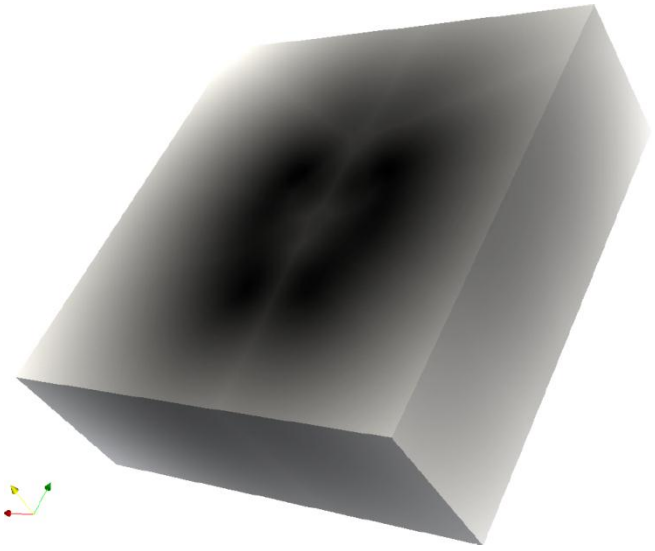


Figure 3. 3D Distance Transform of Figure 2

There is another parallel implementation of Maurer's algorithm, presented in Staubs[2] which features significant speedup for small scale problems and a low number of threads. In this paper, we present a more generalized, scalable, and efficient parallel implementation of Maurer's algorithm. Our implementation is able to provide a mean speedup of 6 times for 8 threads on a machine with 40 cores while Staubs[2] was only able to provide a mean speedup of 3 times with 8 threads on a machine with 8 cores. Also, the speedup data presented in Staubs[2] asymptotically approaches 3 times, while our implementation has a projected asymptote of 20 times for the speedup for large datasets.

2. Our Approach

Our algorithm follows the same dimensionality reduction approach and partial Voronoi diagram generation developed by Maurer[1] to efficiently calculate the EDT for the image. Our algorithm is generalized to compute the EDT for any given number of dimensions. This is done by representing the image as a single array in row-major order. The row-major representation is used instead of column-major representation because the elements are accessed by iterating the index of the current dimension. Row-major order would produce a benefit in access time over column-major order for computing the EDT of a row, while column-major order would produce a benefit for computing the EDT of a column. The array is laid out contiguously in memory which results in better cache performance for iterating over a row or column for row-major or column-major organization, respectively, of the array representation of the binary image.

For our implementation, we are utilizing the POSIX Threads Programming library. The library includes a pthread datatype and a mutex datatype which we used to implement our algorithm. A pthread is the POSIX version of a thread. A thread is an individual collection of instructions to be executed. Without the use of threads, a program runs sequentially, one instruction after another, waiting for each instruction to terminate before beginning the next instruction. With threads, multiple instructions may be executed simultaneously. A mutex is the name given to the semaphore to control access to a common resource. When a thread locks a mutex, then that thread has exclusive control of that resource until the controlling thread releases the resource by unlocking the mutex. If a thread tries to gain control of a mutex that is already owned, or locked, by another thread, then the calling thread waits until the resource becomes available. We also use the thread controls wait, signal, and broadcast. When a thread waits, the thread pauses its execution until it receives a signal. If multiple threads are waiting, then all thread may resume execution through the use of a broadcast.

2.1 Load Balancing

A common problem in parallel algorithms is load balancing. In order to maximize speedup, idle time must be minimized. We have achieved this by utilizing the producer-consumer paradigm in that the main program creates the consumer threads and creates the work for the consumer threads. The total amount

of work is stored in W_d, a one-dimensional array, while the indices of each dimension are stored in $D_{d,w,i}$, a three-dimensional array. These indices are the fixed indices that the consumer thread uses to iterate through the current dimension. The total amount of work for a given dimension can be calculated with a single multiplicative summation if the current dimension is excluded:

$$\left[\prod_{i=0}^{d-1} N_i\right] * \left[\prod_{i=d+1}^{n} N_i\right]$$

Where N_i is the number of rows for dimension i. For each dimension i, there are N_i number of rows. For calculating the EDT, each row of the current dimension must be iterated through for all other dimensions. Consider the example of a 3x4x2 matrix, (width of 3, height of 4, and a depth of 2). For the lowest dimension, we calculate the number of rows that need to be processed by multiplying the number of rows in all other dimensions. So the number of rows that need processing for the lowest dimension would be 4*2=8, and the number of rows that need processing for the next dimension would be 3*2=6, and 3*4=12 for the highest dimension.

Once the work has been generated for the current dimension by the producer, a signal is sent to the consumers which then retrieve work from the front of the queue and compute the EDT by iterating through the indices of the current dimension while keeping the indices of the other dimensions fixed based on the values retrieved from $D_{d,i}$. This approach provides better load-balancing than statically allocating work before processing begins. When a consumer thread checks the queue of work and the queue is empty, the consumer thread goes into a wait state. Once all threads have reached the wait state, a signal is sent to the producer. If the producer has finished generating the work queue for the next dimension, the producer broadcasts to the consumer threads to begin processing the next dimension. A barrier cannot be used because we cannot guarantee that the producer will have produced the work queue necessary for the next dimension by the time the current dimension has finished computing. While the consumer threads process the current dimension, the producer is generating the work queue. This also helps eliminate thread idle time by not waiting until the work queue for all dimensions is generated. This approach is possible because the only task dependency (other than the dependency of the consumer for the producer in generating the work queue) is that the EDT for current dimension is dependent on the previous dimension. Each row in the current dimension is independent of all other rows in the current dimension.

2.2 Dimension Generalization

Once the binary image is represented in row-major form as a single array, the indices for iterating through a given dimension can be computed using the stored size information in N and the current fixed dimensions stored in $D_{i,w}$. We were able to generalize the calculations to compute the indices for a given dimension using a summation of multiplicative summations. These computations are few, even in the case of large dimensions, compared to the task of generating and querying the partial Voronoi diagram and therefore do not have a significant effect on the complexity of the algorithm. Given the current dimension and the indices of the fixed dimension, $D_{d,i}$, the indices of the current dimension are calculated during the construction of the partial Voronoi diagram by

$$index_i = \left[\sum_{k=0}^{d-1}\left(\prod_{j=0}^{k-1} N_j\right) * D_{d,w,k}\right] + \left[\left(\prod_{j=0}^{k-1} N_j\right) * i\right] + \left[\sum_{k=d+1}^{n_d}\left(\prod_{j=0}^{k-1} N_j\right) * D_{d,w,k}\right]$$

This formula works by calculating the offset for each dimension. To calculate the offset for a dimension, the summation of the multiplicative summation of all dimensions is calculated. The multiplicative summation of a dimension is represented by multiplying the size of all lower dimensions to produce an offset. This offset represents the number of voxels that comprise one set of the given dimension. This offset represents the first index of the first set of the dimension. To find the first index of the N-th set of a dimension, the offset is multiplied by N. For our algorithm, the N-th set of the fixed dimensions is given by $D_{d,w,i}$ while the N-th set of the current dimension is given by i. It is important to store these indices for future use while querying the partial Voronoi diagram.

3. Experimental Performance

We have tested our implementation on a single compute node in a cluster containing 52 multi-core compute nodes and 17 Appro GPU nodes. Standard compute nodes are comprised of 28 ivy bridge nodes and 20 sandy bridge compute nodes while the GPU nodes each contain 4 Nvidia Tesla GPU's for a total of 68 GPU's. The compute node that we have access to contains 40 cores and 126 Gb of memory. We

generated cube images with dimensions of 500x500x500, 1000x1000x1000, and 1250x1250x1250 and ran each with a literal interpretation of the algorithm presented in Maurer[1] sequentially and with our implementation using 2, 4, 8, 16, 24, 32, and 40 threads. A cube with dimensions 1250x1250x1250 is used as our largest test case because our machine does not have enough memory to process larger cubes.

The image data that is being processed are cubes, so the total amount of work needed to be done to process an entire cube can be represented by $3*n^2$ where n is the number of rows in each dimension. One unit of work is one row of voxels for a dimension. For our 500x500x500 cube image, there are 750,000 tasks; 1000x1000x1000: 3,000,000 tasks; 1250x1250x1250: 4,687,500 tasks. As the number of threads increases, efficiency decreases. The point at where the efficiency starts decreasing rapidly depends on the problem size. This is due to thread idle time and overhead costs for accessing the mutex to retrieve work. The thread idle time is minimized through dynamic load balancing. Through the use of a mutex, dynamic load balancing is made possible. However, for implementations that use static allocation of tasks, thread idle time would increase due to improper load balancing and there would be minimal overhead costs. For larger problems, using a mutex to minimize thread idle time results in better performance and an overall higher efficiency. Results are depicted in Figure 4 (Speed Up) and Figure 5 (Efficiency). Speed up is defined as the execution time for the sequential algorithm divided by the execution time for the parallel algorithm. Efficiency is defined as speed up divided by the number of threads.

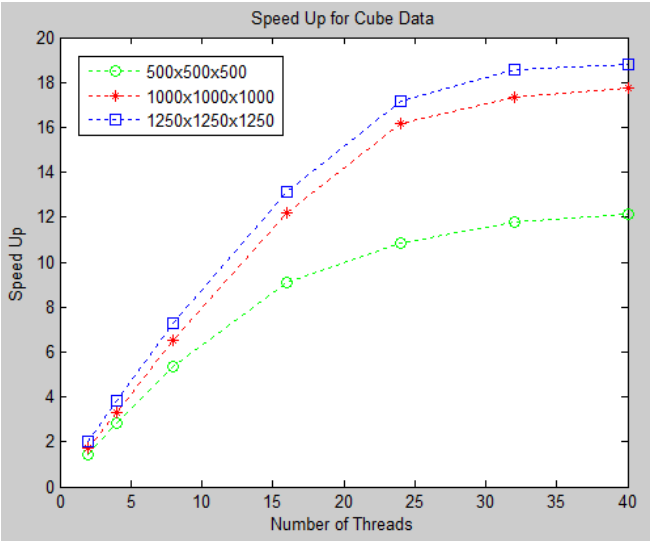


Figure 4. Graph of Speed Up for Cube Data

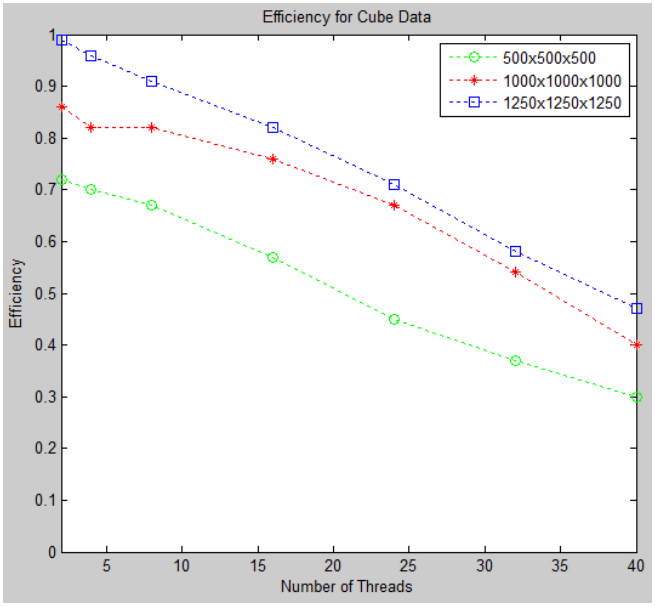


Figure 5. Graph of Efficiency for Cube Data

3.1 Performance Comparisons

The difference between our implementation and the implementation presented in Staubs[2] is that our implementation is more scalable in regards to problem size and number of threads. Our implementation is able to provide a mean speedup of 6 times for 8 threads while Staubs[2] was only able to provide a mean speedup of 3 times with 8 threads on a machine with 4 cores and 2 GB of memory. Also, the speedup data presented in Staubs[2] asymptotically approaches 3 times, while our implementation has a projected asymptote of 20 times for the speedup for large datasets. Using our current machine, we were able to measure mean speedups of 19 times using 40 threads.

3.2 Work In Progress

This implementation is currently still a work in progress as we plan to test our implementation on larger machines and larger datasets. We also plan to further analyze and modify our algorithm to improve the efficiency. Limitations of our implementation include a drop in performance when using a larger number of threads. This is most likely due to the mutex required to access the shared queue of work as the efficiency drops even for larger problem sizes. Another possible adaptation of our implementation that we will examine and evaluate is the master-worker paradigm where the master will communicate to each thread which tasks need to be generated and processed as opposed to our current implementation of the producer-consumer paradigm where the producer

generates the tasks and the consumers process them. The master-worker paradigm will decrease the need for the mutex and alleviate some startup costs on the main program (i.e. the producer in this case). Another possible, but smaller adaptation, would be to create multiple queues of work and multiple mutexes for accessing the work queues. Since the efficiency drops rapidly after 16 threads, one possibility would be to have the number of queues equal to the ceiling of the number of threads divided by 16. Each thread would be designated to an initial queue and when the queue becomes empty, those threads may request work from the other queue(s).

4. Conclusions

Our introduction of a more scalable Parallel Euclidean Distance Transform algorithm implementation will allow for larger datasets to be processed more efficiently and with more processing power. Also, because our algorithm operates on any dimension, we are able to provide an efficient and extendable algorithm for many different image processing applications to utilize without the need to modify our algorithm.

5. Acknowledgements

We thank the Old Dominion University Information Technology Services for making available the computing resources used in our evaluation.

This work was supported in part by the NSF grant CCF-1139864.

6. References

[1] Calvin R. Maurer, Jr., Rensheng Qi, and Vijay Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. IEEE Trans. Pattern Anal. Mach. Intell., 25(2):265–270, 2003. http://dx.doi.org/10.1109/TPAMI.2003.1177156. (document), 1, 2, 3

[2] Staubs R., Fedorov A., Linardakis L., Dunton B., Chrisochoides N. Parallel N-Dimensional Exact Signed Euclidean Distance Transform. 2006 Sep. http://www.insight-journal.org/browse/publication/123.

[3] Talos I-F., Jakab M., Kikinis R., Shenton M.E. SPL-PNL Brain Atlas. SPL-PNL March;

Multi-material Surface Extraction for Sparse Atlas-based Neuroanatomical Representation and Intraoperative Tracking

Tanweer Rashid and Michel A. Audette

Abstract

This paper presents on-going work on deep brain therapy planning, in conjunction with a surgical robotic approach to targeting. The therapy planning and treatment system will exploit a digital deep brain atlas, which will be represented as sparse surface meshes for highly efficient intraoperative registration. While our immediate clinical application is deep-brain stimulation for neurological conditions such as Parkinson's disease (PD), our system will be extensible to emerging treatments, such as genetic therapies. We are developing multi-material, watertight Dual Contouring (DC) for surface mesh extraction from labeled volumes, and 2-simplex meshes for resolution control in the representation of deep-brain structures imbedded in the atlas. A suitably sparse simplex multi-surface representation will then enable, using simplex image forces, the real-time tracking of these subcortical structures, such as the subthalamic nucleus (STN), which is targeted in PD, based on intraoperative, volume-of-interest multi-contrast MRI pulse sequences [1], which exploit gradients coinciding with local variations in iron content in these structures.

While subcortical structures are functionally separate structures, from a neuro-architecture standpoint, they are continuous tracts of white and grey matter, whose intraoperative motion is continuous, not piecewise-rigid or piecewise-affine as is the case for skeletal bones in extremities. As a result, we argue that it is imperative to represent deep-brain structures as having shared boundaries. Consequently, single-boundary contouring, such as classical Marching Cubes or single-surface DC, is not a viable option. This paper will center on the development of multi-material dual contouring, which will serve as a prelude to simplex-based controlled-resolution decimation.

Keywords: Dual contouring, multi-material, simplex mesh, deep brain stimulation, surface extraction

Description of purpose

The purpose of this manuscript is to present a multi-surface triangulated mesh of deep brain structures featuring shared boundaries, which will ultimately be applied to initializing a simplex model by geometric duality. The multi-surface simplex [2] will then be used to produce a controlled-resolution decimation, where simplex image forces will be used to non-rigidly register this model to pre-operative multi-contrast MRI data, as well as perform highly efficient real-time tracking of patient tissue boundaries on the basis of intraoperative MRI. The accuracy of both the pre-operative fitting and intraoperative tracking is essential for the overall accuracy and broad clinical applicability of a MRI-compatible robot designed for delivering therapy to subcortical targets such as the subthalamic nucleus [3].

Method(s)

Dual contouring (DC) is an octree-based method of generating the surface of an implicit function. DC, introduced by Ju et.al in [4] is a dual method in the sense that vertices instead of edges (as in Marching Cubes) are generated to represent surfaces. An attractive feature of DC is that it is capable of reproducing sharp geometric features when hermite data is available. Another advantage of DC is that it can be adapted to

contouring volumes composed of multiple materials, as shown in Figure 1. Our algorithm is an adaptation of a contouring method developed by Feng et.al [5].

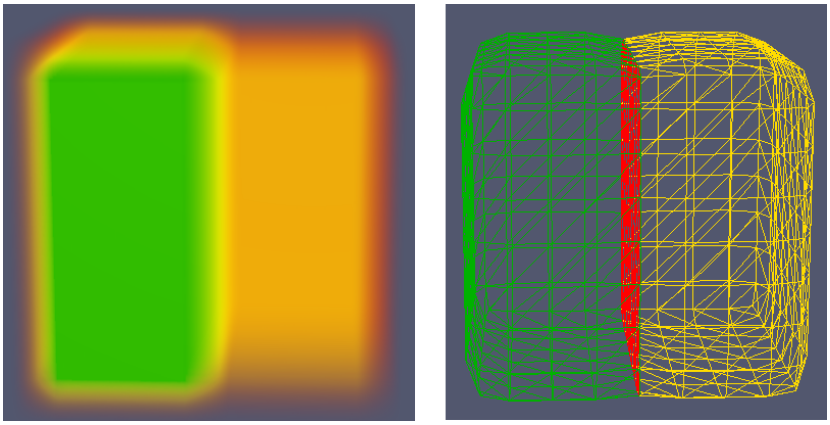


Figure 1: (Left) Volume rendering of a cube with two labels, i.e. composed of two materials. (Right) The mesh generated by DC for the two labels of the cube. Notice the shared boundary colored in red.

The volume of interest is first divided into a uniform grid, and the corners of each cube of the grid is labeled as being inside or outside the surface to be meshed. In Marching Cubes, this labeling consists of simple positive and negative notations. However, for the multi-material DC, the labeling consists of material indices. The octree is created using only those cubes of the grid that have different material indices. Cubes that have the same material indices are considered to be completely within a specified label of the volume, and are thus not considered for surface extraction computations.

$$E(d) = \sum_{i=1}^{n}((d - p_i) \cdot N_i^2 \qquad \text{Eq (1)}$$

Quadratic Error Functions (QEFs) are utilized for vertex computation. Eq 1 shows the typical form of the QEF, where d is the dual vertex to be computed, $E(d)$ is the error of the QEF, p_i is the i^{th} edge intersection, and N_i is the normal at the i^{th} intersection. Solving for the minimum of Eq 1 yields a dual vertex that ideally should lie anywhere inside the grid cube. We refer the reader to [4] for more details.

The original DC paper [4] had two limitations: i) the algorithm generated points outside the grid because of which the surface might have intersecting triangles, and ii) the surface may not necessarily be 2-manifold. Both issues were separately resolved in [6] and [7] by Ju et.al and Schaefer et al. respectively, albeit not for multi-material contouring.

However, to our knowledge, there has not been an implementation of the DC algorithm that generates both intersection-free and 2-manifold multi-material surfaces. In this paper we present an extension of the implementation of Feng et al. [5] by incorporating intersection-free and 2-manifold attributes in conjunction with multi-material contouring.

Simplex meshes were introduced by Delingette in [8]. A k-simplex mesh is one where each vertex is connected to $k + 1$ neighbors. The advantage of simplex meshes is their relatively simple geometric description. Each vertex in a simplex mesh can be described with respect to its three neighbors using two barycentric coordinates and curvature [2]. Simplex meshes use image forces and geometric forces for mesh deformation. Deforming a high resolution mesh (left side of Figure 2) would take more computation time. Gilles in [2] have shown that using a multi-resolution simplex mesh, it is more efficient to initially deform a

coarse (low resolution) mesh, and then increase the resolution during the finer later stages of the deformation process.

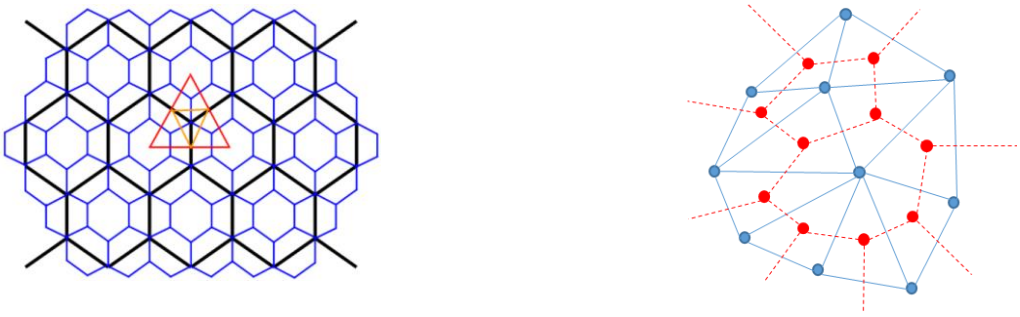
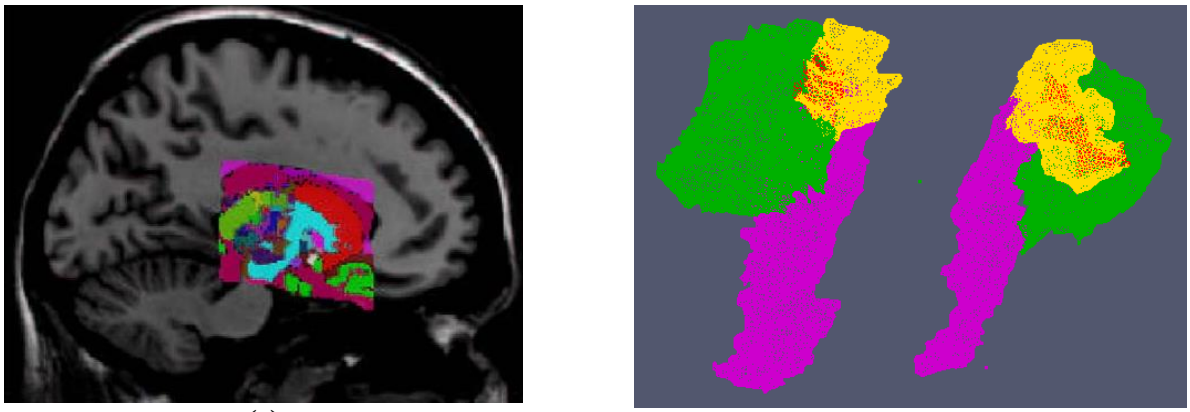


Figure 2: (Left) The multi-resolution simplex scheme in [2]. The black mesh represents the coarse mesh and the blue mesh represents the finer mesh. (Right) A triangular mesh, in blue color, and its dual 2-simplex mesh in red color

Simplex meshes are also simple to construct in the sense that they are dual to triangular meshes. The right side of Figure 2 shows an example of a triangle mesh and its dual simplex mesh.

Results

Figure 2 illustrates preliminary results of our multi-material contouring applied to the digital deep-brain atlas described in [9]. The 2-manifold constraint implementation is still underway.



(a) (b)

Figure 3: Deep brain atlas meshing. (a) Sagittal image of the digital deep-brain atlas. (b) Mesh from DC algorithm. The coloring represents 3 labeled regions in the digital atlas.

Figure 3 shows the results that we have so far. Figure 3(a) shows the Sagittal slice of the region of interest of digital deep brain atlas from [9]. This atlas contains a total of 124 labels. Figure 3(b) shows the results of applying our DC algorithm for 3 labeled regions of the deep brain atlas.

New or Breakthrough work to be presented

Our dual contouring algorithm, whose development is still underway, will be capable of producing surface meshes of a volume composed of multiple materials, or in this case, multiple labeled functional regions. We intend to utilize this algorithm in creating surface meshes for deep brain structures, as an input stage for simplex-based decimation [10], which in turn may be applied to initializing a tetrahedral mesh within one or

more subvolumes. This approach is unique in its suitability to producing a multi-surface mesh representation of deep-brain structures in order to enable real-time tracking of these structures from intraoperative multi-contrast MRI. Without the manifold and watertight qualities, our contouring method will not be applicable to initializing the simplex model by geometric duality, or for use as an input to tetrahedral meshing.

Conclusions

This paper presented the first stage for a surgical robotic approach to targeting in deep brain therapy. We use a dual contouring algorithm capable of producing surface meshes from a volume composed of multiple materials, or in this case, multiple labeled functional regions. We intend to utilize this algorithm in creating surface meshes for deep brain structures, as an input stage for simplex-based decimation [10], which in turn may be applied to initializing a tetrahedral mesh within one or more subvolumes.

Acknowledgements

We would like to acknowledge the support of Dr. Tao Ju and Powei Feng for their support in this research.

References

[1] Y. Xiao, S. Beriault, G. B. Pike, and D. L. Collins, "Multicontrast multiecho FLASH MRI for targeting the subthalamic nucleus," *Magnetic Resonance Imaging*, vol. 30, pp. 627-640, 2012.

[2] B. Gilles, L. Moccozet, and N. Magnenat-Thalmann, "Anatomical modelling of the musculoskeletal system from MRI," presented at the Proceedings of the 9th international conference on Medical Image Computing and Computer-Assisted Intervention - Volume Part I, Copenhagen, Denmark, 2006.

[3] G. Cole, J. Pilitsis, and G. S. Fischer, "Design of a robotic system for mri-guided deep brain stimulation electrode placement," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, 2009, pp. 4450-4456.

[4] T. Ju, F. Losasso, S. Schaefer, and J. Warren, "Dual contouring of hermite data," *ACM Trans. Graph.*, vol. 21, pp. 339-346, 2002.

[5] P. Feng, T. Ju, and J. Warren, "Piecewise tri-linear contouring for multi-material volumes," presented at the Proceedings of the 6th international conference on Advances in Geometric Modeling and Processing, Castro Urdiales, Spain, 2010.

[6] T. Ju and T. Udeshi, "Intersection-free contouring on an octree grid."

[7] S. Schaefer, T. Ju, and J. Warren, "Manifold dual contouring," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, pp. 610-619, 2007.

[8] H. Delingette, "Simplex meshes: a general representation for 3D shape reconstruction," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, 1994, pp. 856-859.

[9] B. G. Chakravarty MM, Hodge CP, Sadikot AF, Collins DL, "The creation of a brain atlas for image guided neurosurgery using serial histological data," *NeuroImage* pp. 30: 359 – 376, 2006.

[10] D. H. Audette MA 07, Fuchs A., Astley O. and Chinzei K., "Topologically Faithful, Tissue-guided, Spatially Varying Meshing Strategy for Computing Patient-specific Head Models for Endoscopic Pituitary Surgery Simulation," *Journal of Computer Aided Surgery*, pp. 12(1): 43–52, Jan. 2007.

Modeling of Cranial Nerve Using 1-Simplex Mesh

Sharmin Sultana, Michel A. Audette

Extended Abstract – Segmentation of tubular structure from 3D medical images such as CT or MRI is of vital interest for medical applications like diagnosis and surgical planning. When it comes the question of segmenting tubular structure like cranial nerve, it becomes more challenging. We have represented here the centerlines of cranial nerves using 1-simplex mesh. Later on, from these centerlines we are planning to estimate the radii using image intensity information.

A k-simple mesh is a k-manifold discrete mesh where each vertex has a constant connectivity i.e. each vertex has exactly k+1 distinct neighbors [3]. Neighboring vertices are connected by edges, edge closed successions form faces and faces closed successions form cells. Based on the connectivity *k*, simplex meshes can represent various object such as curves ($k = 1$), surfaces ($k = 2$) or volumes ($k = 3$). A *k*-simplex mesh is topological dual to a *k*-solid mesh. For example, 1-simplex mesh is dual to polyline, 2-simplex is dual to triangular mesh and 3-simplex is dual to tetrahedral mesh.

We are using 1-simplex mesh to represent the centerline

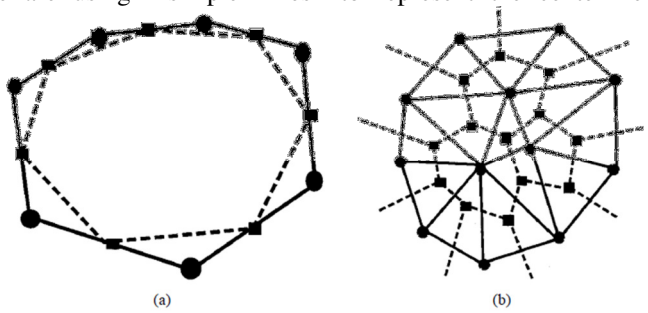


Figure 1: a) 1-simplex mesh with its dual polyline; b) 2-simplex mesh with dual triangulation. Simplex meshes are drawn with dashed line and corresponding dual are drawn with solid line.

of the cranial nerves. The geometry of 1-simplex is defined through two angular parameters and two matric parameters. These parameters define the local shape around a given vertex. The simplex angle controls the height of a vertex with respect to the tangent plane and the metric parameters control the vertex position in the tangent plane with respect to its two neighbors.

Let P_i be a vertex of a planner 1-simplex mesh and F_i is the projection of P_i onto the line $[P_{i-1}P_{i+1}]$. F_i can be represented as a weighted sum of two neighboring points:

$$F_i = \epsilon_{1i} P_{i-1} + \epsilon_{2i} P_{i+1}$$

$$\epsilon_{1i} + \epsilon_{2i} = 1$$

The simplex angle, φ_i is the oriented angle between the two adjacent segment $[P_{i-1}P_i]$ and $[P_iP_{i+1}]$ as shown in Figure 2.

Another angle, ψ_i is defines as the angle between the normal vector, *n* and the vector, *u*.

A 1-simplex point P_i can be fully defines as:

$$P_i = \epsilon_{1i} P_{i-1} + \epsilon_{2i} P_{i+1} + h(\varphi_i) n(\psi_i)$$

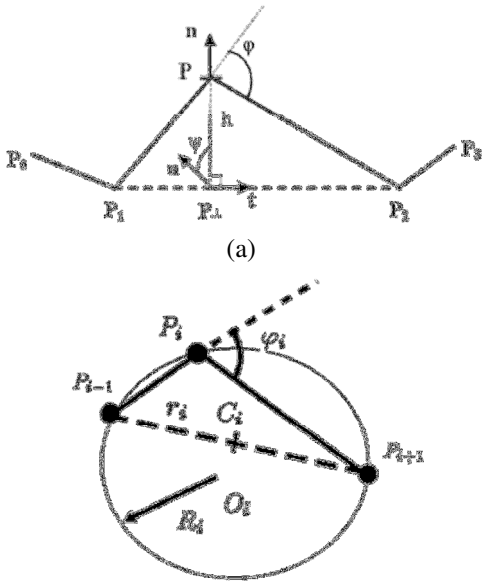


Figure 2: Geometry of 1-simplex mesh. a) Tangent and Normal plane around a vertex P_i [2]. b) Simplex angle in its circumscribed circle.

To assure some level of shape continuity, an internal force, F_{int} is applied to each vertex of the mesh. F_{int} is composed of tangential force and normal force:

$$F_{int} = F_{Tangent} + F_{Normal}$$

Tangential force, $F_{Tangent}$ controls the position of the vertex with respect to two neighbors in the tangent plane. Tangential force acting on a vertex P_i can be expressed as :

$$F_{Tangent} = (\check{\epsilon}_{1i} - \epsilon_{1i}) + (\check{\epsilon}_{2i} - \epsilon_{2i})$$

Where $\check{\epsilon}_{1i}$ and $\check{\epsilon}_{2i}$ are reference metric parameters.
Normal force, F_{Normal} constrains geometric continuity. It can be written as:

$$F_{Normal} = \big(L(r_i, d_i, \breve{\varphi}_i)\cos(\breve{\psi}_i) - L(r_i, d_i, \varphi_i)\cos(\psi_i)\big)r_i + \big(L(r_i, d_i, \breve{\varphi}_i)\sin(\breve{\psi}_i) - L(r_i, d_i, \varphi_i)\sin(\psi_i)\big)t_i \wedge r_i$$

Now, if $\breve{\varphi}_i = \varphi_i$ and $\breve{\psi}_i = \psi_i$ then the normal force is null and a C^0 continuity can be achieved. If we set $\breve{\varphi}_i = \breve{\psi}_i = 0$, then a C^1 continuity can be achieved. A C^2 constrain can be maintained by setting $\breve{\psi}_i = 0$ and $\breve{\varphi}_i = \frac{\varphi_{i-1}+\varphi_i+\varphi_{i+1}}{3}$. Figure 3 shows 1-simplex curve representation of synthetic points.

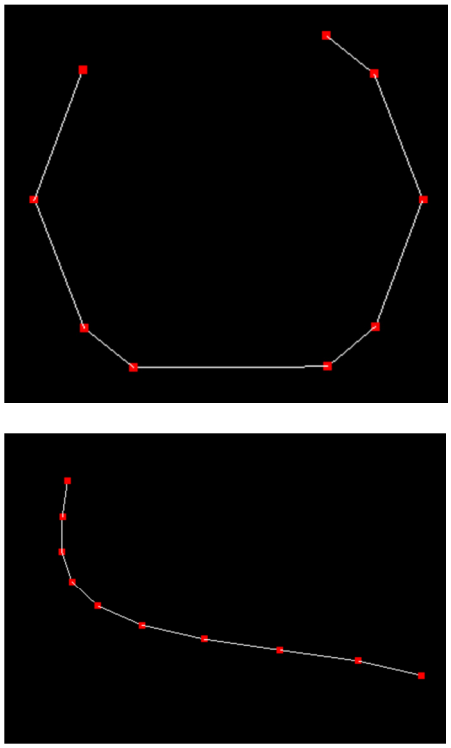


Figure 3: Validation of 1-simplex mesh using synthetic data.

In this method, we are planning to develop a minimally supervised segmentation technique that requires a very little amount of user interactions. The algorithm would only require two end-points of each cranial nerve - one with brain stem and one at the foreman. From this two seed points, centerline of the cranial nerve will be calculated and then using image intensity information [1] and shape prior the 3D model of the cranial nerve will be constructed.

References

[1] Stephen R. Aylward, "Initialization, Noise, Singularities, and Scale in Height Ridge Traversal for Tubular Object Centerline Extraction" IEEE Transaction on Medical Imaging, VOL. 21, NO. 2, Feb 2002.

[2] Benjamin Gilles, Nadia Magnenat-Thalmann , "Musculoskeletal MRI segmentation using multi-resolution simplex meshes with medial representations", Medical Image Analysis 14 (2010) 291–302.

[3] Herve Delingette, "General Object Reconstruction based on Simplex Meshes" , International Journal of Computer vision, 32, 111-142

Gaming & Virtual Reality Application

VMASC Track Chair: Dr. Hector Garcia
MSVE Track Chair: Dr. Yuzhong Shen

Applying Discrete Laplace-Beltrami Operator to Mesh Color Sharpening
Author(s): Zinat Afrose, and Dr.Yuzhong Shen

Effect of Music's Genre on the Height of Peak Sound Waves
Author(s): Ruofan Shen, and Monika Getsova

Applying Discrete Laplace-Beltrami Operator to Mesh Color Sharpening

Zinat Afrose and Yuzhong Shen
Department of Modeling, Simulation, and Visualization Engineering
Old Dominion University
zafro001@odu.edu, yshen@odu.edu

Keywords: Laplace-Beltrami Operator, Colored Mesh, Sharpening.

Abstract

This paper presents a new method for mesh color sharpening using the discrete Laplace-Beltrami operator, which is approximation of second order derivative for irregular 3D meshes. The one-ring neighborhood is utilized to compute the Laplace-Beltrami operator. The color for each vertex is updated by adding the Laplace-Beltrami operator of the vertex color weighted by a factor to its original value. Experimental results demonstrated the effectiveness of the proposed algorithm.

1. INTRODUCTION

Three-dimensional (3D) meshes are widely used in many fields and applications, such as computer graphics, games, animation films, virtual reality, and medical visualizations. 3D meshes are usually generated using one of two methods: 1) artists create the meshes from scratch with 3D modeling software, such as Autodesk Maya and Google SketchUp; and 2) the meshes are created by scanning real 3D objects. The second method is becoming more popular because of the increased precision and processing power of 3D scanners with decreasing cost at the same time.

Microsoft Kinect is a motion sensing device used by Microsoft Xbox 360 and Xbox One game consoles and Windows PCs. The Kinect for Windows sensor contains three components that are related to motion sensing: an RGB camera, an infrared (IR) emitter and an IR depth sensor. The RGB camera can capture color images with a resolution of 1280 × 960 at 12 frames per second or 640 × 480 at 30 frames per second. The IR emitter emits infrared light beams and the IR depth sensor receives the IR beams reflected by the player(s) or environment, from which the depth or distance between an object and the IR depth sensor is computed. After its release, there have been many attempts to use Kinect as a 3D scanner, as it is very economical with a price of $150, compared with scanners with typical prices of thousands or tens of thousands dollars. One of the best available applications that utilize Kinect is ReconstructMe [1]. Combined with the continually decreasing cost of 3D printers, low-cost 3D scanners such as Kinect will have a bright future for 3D modeling at home.

Various methods have been proposed to improve the quality of the mesh generated by 3D scanners, such as surface smoothing, which removes geometrical noise in the surface mesh. ReconstructMe creates color meshes with each vertex of the mesh containing position, normal, and color information. To improve the quality of color meshes, two approaches can be utilized: geometrical processing and color (appearance) processing. Geometrical processing changes each vertex's position while keeping its color information intact; on the contrary, color processing changes each vertex's color while keeping its position (or the object shape) intact. This paper proposes a new method to enhance the visual appearance of the mesh, namely, a mesh color sharpening method using the discrete Laplace-Beltrami operator.

2. METHOD

Sharpening is commonly used in image processing to highlight transitions (or edges) in intensity. One important approach for image sharpening is the Laplacian, which is a second order derivate. Because the pixels in an image are arranged in a rectangular grid, the discretization of Laplacian is straightforward [2] and is computed as second order differences along horizontal and vertical directions. For most 3D meshes, no such rectangular grids exist, so imaging sharpening methods cannot be directly applied to mesh color sharpening.

The Laplace-Beltrami operator is the generalization of Laplacian to functions defined on surfaces in Euclidean space. Let f be a C^2 real valued function defined on a differentiable manifold M with Riemannian metric. The Laplace-Beltrami operator is [3]

$$\Delta f := \operatorname{div}(\operatorname{grad} f), \qquad (1)$$

where grad and div are the gradient and divergence on the manifold M [3].

For discrete meshes, the function f on a triangular mesh T is defined by linearly interpolating the values $f(p_i)$ of f at the vertices of T. This is done by choosing a base of piecewise linear *hat-functions* φ_i, with value 1 at vertex p_i and 0 at all the other vertices[4].

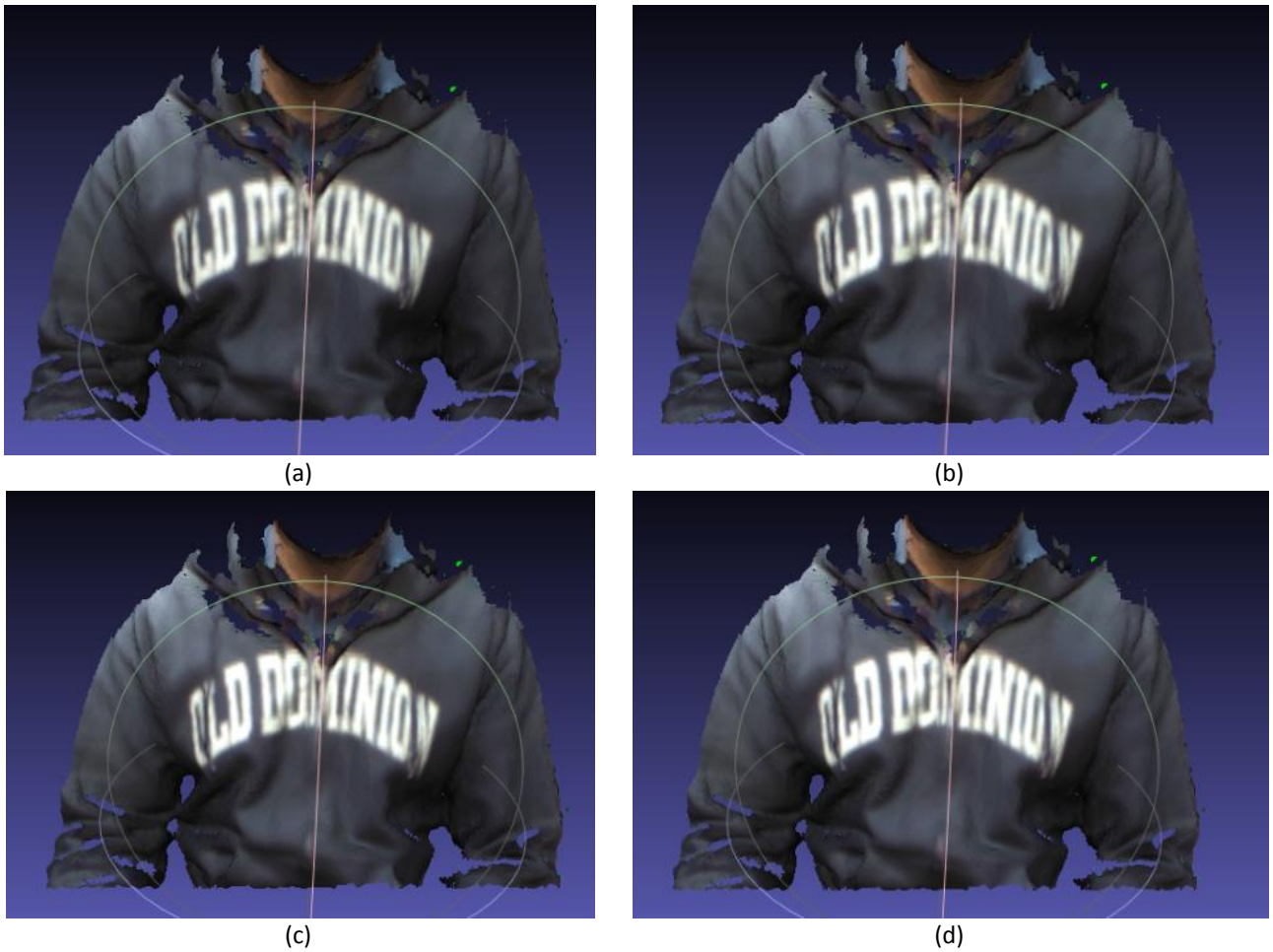


Figure 1: Mesh color sharpening (a) Original scanned mesh. Sharpening results at (b) 1st iteration, (c) 2nd iteration, and (d) 4th iteration.

Then f is given as

$$f = \sum_{i=1}^{n} f(p_i)\varphi_i. \quad (2)$$

Discrete Laplace-Beltrami operators are usually represented as [5]

$$\Delta f(p_i) = \frac{1}{di} \sum_{j \in N(i)} w_{ij}[f(p_i) - f(p_j)], \quad (3)$$

where $N(i)$ denotes the index set of the 1-*ring* neighborhood of the vertex p_i, i.e., the indices of all neighbors connected to p_i by an edge. The mass d_i is associated to a vertex i and the w_{ij} are the symmetric edge weights. This paper utilizes constant masses (i.e., d_i =1) and weights computed as follows,

$$w_{ij} = \frac{\cot(\alpha_{ij}) + \cot(\beta_{ij})}{2}, \quad (4)$$

where α_{ij} and β_{ij} denote the two angles opposite to the edge (i, j).

Each vertex color contains 3 channels: red, green, and blue. Each color component is processed separately. The Laplace-Beltrami operator is calculated for each color component of a vertex and then that color component is updated by adding its Laplace-Beltrami operator weighted by a factor to its original value. This operation is repeated for all color components of all vertices.

3. RESULTS

We applied the proposed mesh color sharpening algorithm to several 3D models we captures using Kinect. Some results are shown in Figure 1. As can be seen, the sharpening operation improved the visual quality of the mesh, which appears crisper after the sharpening operation.

In the future, other discretization of Laplace-Beltrami operator will be implemented for sharpening the mesh color. We also plan to port other image processing algorithms to 3D meshes.

References

[1] ReconstructMe. (2014). *ReconstructMe*. Available: http://reconstructme.net/

[2] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1992.

[3] I. Chavel, *EigenValues in Riemannian Geometry*. Orlando, FL: Academic Press, 1984.

[4] M. Reuter, S. Biasotti, D. Giorgi, G. Patane, and M. Spagnuolo, "Discrete Laplace-Beltrami Operators for Shape Analysis and Segmentation," in *IEEE International Conference on Shape Modelling and Applications*, 2009, pp. 381-390.

[5] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr, *Discrete differential-geometry operators for triangulated 2-manifolds*: Springer Berlin Heidelberg, 2003.

Effect of Music's Genre on the Height of Peak Sound Waves

Ruofan Shen and Monika Getsova
Mathematics & Science Academy
Ocean Lakes High School
ruofanshen@yahoo.com, monikagetsova@yahoo.com

Keywords: Music, genre, sound wave, oscillations

Abstract

This extended abstract is based on a symposium project that was completed as part of the curriculum requirement of Mathematics & Science Academy at Ocean Lakes High School. This project was created in order to identify patterns in the oscillations of sound waves emitted by several genres of music, which included classical, country, pop, and dubstep. Five songs from each of the genres were played for three minutes and the highest and lowest wave peak shown on an oscilloscope screen were recorded at intervals of thirty seconds. The results were measured using an oscilloscope system specially constructed just for this experiment.

1. METHODS

Oscilloscopes are used in a wide array of scientific fields including engineering, telecommunications, and medicine. One of the most common uses of the oscilloscope is as a heart rate monitor, as shown by the continuous fluctuating green line [1]. Other uses in medicine include the measuring of brain activity through brain waves and detecting waves and energy for treating cancer. An engineer may use an oscilloscope to measure engine vibrations or detect a faulty circuit [2]. Although there are new types of digital and far more advanced oscilloscopes in the scientific community, they are all still derived from the basic idea of transmitting and measuring wavelengths and frequencies.

The goal of this project was to gain knowledge on the properties of waves, what specific wave qualities define a genre of music, and how an oscilloscope works. This could've been done with a premade oscilloscope kit, but this was not the purpose. Due to this, a majority of the background research was focused on determining how to convert an old television into a relatively accurate oscilloscope. To do so, one must understand what an oscilloscope is, and how this specific one works. An oscilloscope is used to convert electricity into light; the light is seen on a screen in the form of waves. It is a useful tool because it shows the amount of voltage entering the device at any given time and shows how the voltage has increased or decreased. The main parts of any oscilloscope are the cathode ray tube and the electron deflection plates. The plates that control the vertical aspect of the display are connected to the source of the voltage and the plates responsible for the horizontal aspect are attached to a clock mechanism [1]. An oscilloscope display essentially shows a graph of two variables, time and voltage, residing on the x and y-axis, respectively. In this experiment and in the music, the oscilloscope has a different purpose. The time and voltage variables can also be interpreted as frequency and pitch of sound waves that have been converted into an electric form of oscillation [2].

While doing this research, it became quite evident that a TV could be turned into an oscilloscope, since it possesses all of the necessary parts. A traditional oscilloscope requires a cathode ray tube; therefore, the TV being used had to be a relatively old model. The electron gun is the part of the TV that sticks out towards the person when the back is opened, the deflector coils are a set of wires wrapped around the edge of the gun and attached to the circuit board, and the cathode ray tube is the large cone that connects these parts to the screen. One must then separate the coils and identify which one serves as the vertical deflection coil and vice versa. This can be done with trial and error by switching out pairs of wires and plugging the TV in, such as using different combinations of red, green, yellow, and blue wires. From this point on, one must simply arrange the coils as the plates would be arranged on an actual oscilloscope and attach a voltage source, such as an amplified a music player. Display on the screen may not show any change at all because the TV is not sensitive enough to pick up the small amounts of voltage from the music device, so an amplifier is needed.

2. RESULTS

Five songs from each of the genres were played for three minutes and the highest and lowest wave peak shown on an oscilloscope screen were recorded every thirty seconds. The results were measured using a system specified just for this experiment. The X-axis of the oscilloscope was the zero mark, all waves above it were marked with a positive value in centimeters and all waves below it were negative. The hypothesis was that the genres that are characteristic of including more bass, such as dubstep, would show more dramatic shifts from positive to negative in their analog waveforms (Figures 1 and 2). The

hypothesis turned out to be partially supported by the data because the genre with the most dramatic shifts was actually pop (5.0 cm to -5.0 cm); however, pop music includes large amounts of bass or percussion as well (Figure 3). Country music had more subtle shifts in wave height, as the highest was from 3.1 cm to 5.0 cm (Figure 4). As implied, classical music presented the least amount of shifting, with the highest being from 2.5 cm to -2.3 cm (Figure 5).

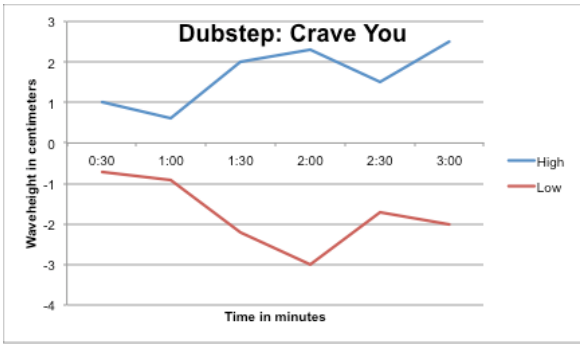


Figure 1. Dubstep song "Crave You" wave heights in cm

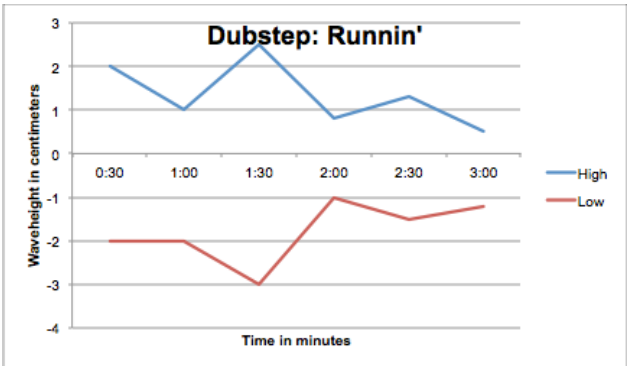


Figure 2. Dubstep song "Runnin'" wave heights in cm

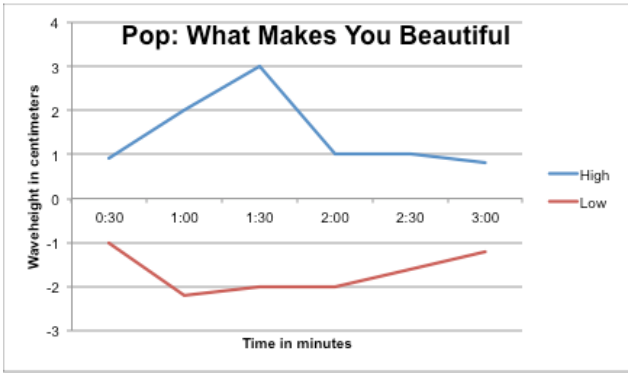


Figure 3. Pop song "What Makes You Beautiful" wave heights in cm

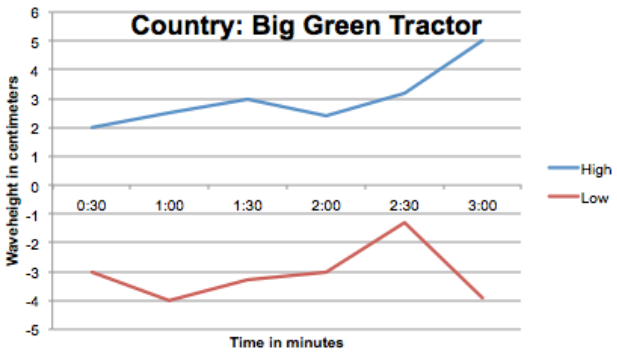


Figure 4. Country song "Big Green Tractor" wave heights in cm

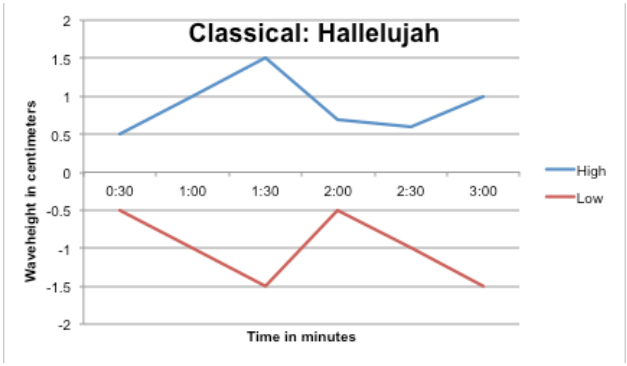


Figure 5. Classical song "Hallelujah" wave heights in cm

3. REFERENCES

[1] UXL Science, Oscilloscope, Science in Context, Retrieved from http://ic.galegroup.com/ic/scic/ReferenceDetailsPage/ReferenceDetailsWindow?query=&prodId=SCIC&displayGroupName=Reference&limiter=&source=&disableHighlighting=false&displayGroups=&sortBy=&search_within_results=&action=2&catId=&activityType=&documentId=GALE%7CCV2646000723&userGroupName=va_s_128_0960&jsid=2c25bec6fc76466946005bf2e7161af3, 2014.

[2] University of Evansville, The Oscilloscope, Retrieved from http://uenics.evansville.edu/~amr63/equipment/scope/oscilloscope.html, 2013.

[3] Tektronix, Oscilloscope Fundamentals, Retrieved from http://circuitslab.case.edu/manuals/Oscilloscope_Fundamentals_-_Tektronix.pdf

Agent Based Modeling & Simulation

VMASC Track Chair: Dr. Andrew Collins
MSVE Track Chair: Dr. Ahmed Noor

Infection Dynamics on a Risk-Benefit Evolving Social Network
Author(s): Shadrack Antwi, and Leah Shaw

Beyond Opinion Polls: Predicting Outcomes of Independence Referenda
Author(s): Jan Nalaskowski

Agent-Based Simulation Event Execution Architecture for Improved Performance and Scalability
Author(s): Jesse Cladwell, Tyrell Gardner, and Dr. Jim Leathrum

Modeling Effectiveness of Tick Control by a Species that Exhibits Predator-prey Role Reversal
Author(s): Alexis White, Robyn Nadolny, Carrie Eaton, and Holly Gaff

Popularity or Proclivity? Revisiting Agent Heterogeneity in Network Formation
Author(s): Xiaotian Wang, and Dr. Andrew Collins

Infection dynamics on a risk-benefit evolving social network

Shadrack Antwi Leah Shaw
The College of William and Mary, Virginia

February 1, 2014

Abstract

Social dynamics models have been used to study how behavior affects the dynamics of an infection in a population[**1**]. There are two main directions of these investigations. The first is agent-based aggregate models which look at the behavior of a roughly uniform mixture of individuals[**2,3**]. The other uses complex networks where nodes or vertexes represent individuals and links or edges are the relationships between them[**4**]. Static networks were used by many authors to study disease transmission when relationships between individuals are fixed (e.g., [**5,6**]). A more realistic extension is adaptive networks, where individuals are capable of changing their social connections in response to the presence of a disease[**7,8**]. Various means of changing the social circles of nodes in response to the presence of disease have been implemented in the study of how behavior modification can affect disease propagation[**1,4**].

In their discussion section, [**8**] raised the possibility of modeling diseases such as HIV where people may not necessarily know each other's infection state. The present paper[**9**] aims to study this scenario. We are motivated by HIV. Individuals with more sexual partners are more likely to acquire HIV[**10**]. Thus where HIV status is unknown people may try to avoid those with excessively high number of partners. We build a social network model for individuals making and breaking sexual connections based on each other's desirability as a partner. An individual's desirability here is understood as the benefit others derive in a relationship with it compared with its risk of transmitting the disease to its partners. Individuals therefore seek a balance between finding more partners and cutting off high risk partners to reduce the chance of infection.

We define a stochastic network model with formation and breaking of links representing changes in sexual contacts. Each node has an intrinsic benefit that its neighbors derive from connecting to it. We assume that for a particular node the same benefit value is seen by all potential neighbors. Increasing its number of partners increases the total benefit that a node derives, but this also increases its risk of infection. A node's infection status is unknown to others. However, nodes with higher degree (number of connections) are assumed more likely to be infected[**11**]. We define a payoff function that captures a node's desirability as a connection, which increases with increasing benefit and decreases with increasing degree. Each node's goal, then, is to form links to high payoff nodes and break from those with low payoff.

We implement the social dynamics in the absence of disease to study how a population that is oblivious to the presence of an infection behaves. This also applies when there is no infection in the population. We determine the level of activity by characterizing the network connectivity. The interaction between human behavior and the disease is obtained by modeling a SI (susceptible-infected) epidemic concurrent to the social dynamics on the complex network. We determine dependence of network connectivity and infection levels on parameters. We also obtain steady state degree distributions and infection thresholds.

We actualized the network and epidemic dynamics by Monte Carlo simulations. We also derive analytic approximations of the system behavior by a heterogeneous mean field (HMF) approach, writing differential equations for the number of nodes with each degree[**5**]. The HMF theory was in good agreement with the network simulation.

When the entire population is informed by health services of the disease's presence and also the general epidemic level, individuals may take further action to avoid the disease by cutting down on their number of partners or activity level[**1,2,3,12**]. We extend the model to consider the instance where infection levels across the entire network change the risk aversion of individual nodes. This case is an adaptive network, where network state affects the state of nodes; the state of nodes then feeds back to affect the network structure and state[**13**].

The model considered so far has all nodes having the same benefit value. However, in real life individuals may differ in their intrinsic benefits. These differences may affect how they fare

in their social interactions. Since sexual contacts are a primary means of HIV transmission, any intrinsic differences among individuals which cause them to achieve different number of partners could result in their having different chances of infection. We study the effects of such differences by modeling the social and infection dynamics for a (non-uniform) population with two benefit values. We determine how disease and mortality differs for the two types of nodes. We further study how changes in the make-up of newborns, in terms of their benefit values, would affect the epidemic and the make-up of the entire population.

References

[1] S. Funk, M. Salathe, V. A. A. Jansen, *Modelling the influence of human behavior on the spread of infectious disease: a review* , J. R. Soc. Interface (2010) 7, 1247-1256

[2] J. Epstein, J. Parker, D. Cummings, R. Hammond, *Coupled contagion dynamics of fear and disease: mathematical and computational explorations*, PLoS ONE (2008) 3(12) e33955

[3] T. Reluga, *Game theory of social distancing in response to an epidemic*, PLoS Computational Biology (2010) 6 e1000793

[4] L. Danon, A. Ford, T. House, C. Jewell, C. Keeling, M. Roberts, G. Ross, M. Vernon, *Networks and Epidemiology of Disease*, Interdisc. Persp. on Infectious Disease, 2011 (2011)

[5] R. Pastor-Satorras, A. Vespignani, *Epidemic spreading in scale-free networks*, Phys. Rev. Lett. 86, 3200 (2001)

[6] R. M. May, A.L. Lloyd, *Infection dynamics on scale-free networks*, Phys. Rev. E 64, 066112 (2001)

[7] T. Gross, C.J.D. D'Lima, B. Blasius, *Epidemic dynamics on adaptive networks*, Phys. Rev. Lett. 96, 208701 (2006)

[8] L. Shaw, I. Schwartz, *Fluctuating epidemics on adaptive networks*, Phys. Rev. E 77, 066101 (2008)

[9] S. Antwi, L. Shaw, *Epidemic dynamics on a payoff-dependent evolving social network*, In preparation

[10] B. Varghese, J.E. Maher, T.A. Peterman, B.M. Branson, R.W. Steketee, *Reducing the risk of sexual HIV transmission:quantifying the per-act risk fo HIV on the choice of partner, sex, act, and condom use*, Sex. Transm. Dis. 29(1):38-43 (2002)

[11] R.M. Christley, G.L. Pinchbeck, R.G. Bowers, D. Clancy, N.P. French, R. Bennett, J. Turner, *Infection in social networks: using network analysis to identify high-risk individuals*, Am. J. Epidemiology (2005) 162(10): 1024-1031

[12] E. Fenichel, C. Castillo-Chavez, M. Ceddia, G. Chowell, P. Gonzalez Parra, G. Hickling, G. Holloway, R. Horan, B. Morin, C. Perrings, M. Springborn, L. Velasquez, C. Villalobos, *Adaptive human bahavior in epidemilogical models*, PNAS (2011) 108(15), 6306-6311

[13] T. Gross, B. Blasius, *Adaptive Coevolutionary networks: a review*, J. R. Soc. Interface (2008) 5, 259-271

Beyond Opinion Polls: Predicting Outcomes of Independence Referenda

Jan Nalaskowski
Graduate Program in International Studies
Old Dominion University, Norfolk, VA 23529, USA
jnalasko@odu.edu

Keywords: independence referenda, political campaigns, Social Judgment Theory, Scotland

Abstract

This paper introduces modeling assessment of preparations for Scottish independence referendum. Taking into account data retrieved from opinion polls and general elections it proposes four classes of agents communicating with each other within the framework of the Social Judgment Theory and during simulation consisting of 1460 days. The agent-based model aims to test various statements formulated by commentators of the referendum event. There are many opinions on how both pro and anti-independence campaigns should look like, what parts of electorate should they focus on and what are the real chances of independence to be chosen, facing relatively constant lack of support for the case indicated by opinion polls. The main findings are that under current distribution of agents the intensity and scope of pro-independence campaign as well as the choice of its recipients do not really matter, as there is virtually no chance for independence to be chosen.

1. INTRODUCTION

Recent developments on the Scottish political scene have revealed discrepancy between fixed and relatively low popular support for independence on the one hand, and significant electoral support obtained by the Scottish National Party in the last elections on the other. Scoring majority of seats in the parliament enabled the party to announce referendum on independence to be held in September 2014. This decision is controversial since opinion polls reveal consistent lack of support for Scottish sovereignty. Does the party therefore count on pro-independence campaign to secure required support? As history shows, many referenda outcomes proved to be unexpected so there is surely a temptation to believe that no matter how low a support for independence seems to be, a skilfully shaped campaign can help in winning the case.

The model presented here uses the Social Judgment Theory with its elegant framework to assess opinion changes among interacting population. The theory is formalized with agent-based modeling and simulations are run to obtain meaningful conclusions.

The paper is divided into five sections. First, the simuland is explained and main assumptions retrieved from the research are pointed out. Second, theoretical underpinnings are reviewed. The model is based on the Social Judgment Theory, rationale obtained from theories on referenda and the real-world data. Third, the model itself, agents and logic of simulation are explained. Fourth, the results of multiple simulations are summarized. This section also points out important validation and verification techniques which were used. Finally, the last section introduces conclusions together with limitations and further research pointing out the way to improve the model in future work.

2. SIMULAND

Scotland's parliamentary elections of 2007 resulted in victory of Scottish National Party – for the first time in history of devolution. This outcome became particularly significant for supporters of seceding from the Great Britain. An "uncertain path towards independence" was launched [Wintour 2007]. This uncertainty has been reinforced by the fact that majority of Scots consequently don't want sovereignty, even though Scottish National Party managed to magnify its number of votes in subsequent parliamentary elections in 2011. The reason why the party won was surely not its advocacy for independence. It was rather about attracting working class votes and successful achievements including personal care and tuition fees, while in contrast Labour party was associated with economic crisis under prime ministership of Tony Blair and Gordon Brown [Wray 2011].

According to current opinion polls, only about 55 percent of Scottish National Party electorate would choose independence in a referendum [Carrell 2013a]. The long-term data trend indicates that supporting Scottish National Party in parliamentary elections does not converge with increase of independence support and this is most probably how the future developments will look like [Dinwoodie 2011].

Despite all odds, the ruling party has decided to introduce referendum on independence, which is scheduled for September 2014. Up to date, there were two referendums held in Scotland, both on the devolution-related matters. The one in 1979 was inconclusive because of additional constrain requiring that the outcome represented 40 percent of the total Scottish electorate, rather than simple majority of those who voted. The result was a slight majority supporting the step towards devolution which however did not meet required 40 percent so the change was not implemented. The second devolution referendum took place in 1997 and majority of voters - 74 percent - opted for devolution. Prior opinion polls indicated between 50 and 60 percent for those in favor and 20 – 30 percent for those against [Denver et al. 2000: 123-124]. For 1979 referendum opinion polls showed 35-49 percent in favor and 33 percent against [Taylor]. Though opinion polls were not very precise in predicting outcomes, they correctly pointed out which side would gain majority of support. This fact puts in question the success of independence option in upcoming referendum.

Both pro-independence and pro-union camps have launched major campaigns to support their respective cases. The Scottish National Party has given itself some time for working on campaign and appealing to both swing voters and to those undecided, counting on their relatively high number. The party particularly wants to target working class urban voters in central Scotland, who

are most likely to support independence but also least likely to vote [Carrell 2013b]. The final campaigns of both camps will be launched in May 2014 and intensification of actions is likely to appear around then [Carrell 2013c].

There are many opinions on how campaigns should look like. According to the polling expert Nate Silver, opinions data is quite definitive and there is virtually no chance for independence to be chosen [Higgins 2013]. On the other hand, according to the polls from October 2013, the proportion of those undecided has reached 31 percent, suggesting that current campaigns fail to provide enough information for people to make up their minds [McLaughlin 2013]. At that time, at least 14 percent of respondents confirmed the lack of sufficient information while 44 percent said that they were poorly informed [Peterkin 2013]. These outcomes suggest that there is significant number of Scotsmen who could potentially become recipients of campaigns activities.

The other question is how intense should these activities be. Many opinions indicate that there is no linear relation between pervasiveness of campaign and convincing voters to choose particular option. People usually distance themselves from extreme options. "For the Nationalists, the lessons are not to obsess about independence as an end in and of itself, but to relate it to self-government as a means of making a significant difference to material and social concerns" [McCrone 2012: 76]. Even though there are voices that both "yes" and "no" campaigns should be more aggressive [Carrell 2012], there are also those who are worried about toxicity of actions. Scottish author Denise Mina recently appeared on radio program to discuss referendum issues, finding herself trapped between two extreme pro and anti-independence advocates. As she concluded, she was not only undecided but also "sick to death" by the nature of debate. Mina added that it is intelligent and public discourse that really should play the leading role in campaigns [Mina 2013].

3. THEORETICAL ASSUMPTIONS

3.1. Nature of Referenda

According to Samantha Laycock, the difference between general elections and referenda is that in the former the question of policies is only one among many considerations while in the latter political parties are of secondary importance, as the decision is made about particular provision to be implemented. Usually referenda are considered less important than national elections [Laycock 2013: 237-238].

It should however be noted that independence referenda may be an exceptional case because of their emotional implications. To date, 49 independence referenda have taken place. After 1945, 62 percent of them resulted in "yes" option. In general, the turnout has been higher than in regular elections, with average mean of 79 percent [Qvortrup 2013: 4-5].

When it comes to the impact of campaigns on referenda outcomes, political scientists are inconclusive. Recently much has been said about the positive relation between these two, but formal analyses show no statistically significant difference between referenda campaigns and general elections campaigns [Laycock 2013: 243-247]. Party affiliation has smaller impact on referenda outcomes but it doesn't mean that it has no influence at all. In Scottish devolution referendum of 1997, Labour party, Scottish National Party and Liberal Democrats maintained unified front and campaigned together.

The conclusion is that in the case of Scottish independence referendum one can expect quite significant turnout. While party identification is probably not as important as in general elections, some positive influence is expected to emerge. The impact of campaign is apparently not strictly linear and both positive and negative relations can be observed.

3.2. Social Judgment Theory

The Social Judgment Theory offers an elegant and widely accepted way of assessing update of opinions that people express after being confronted with views advocated by their communication partners [Sherif and Hovland 1961]. According to the theory, people tend to be positive towards opinions close to their own, while distancing themselves from those remote. This dynamics of "boomerang effect" [Chau et al. 2013: 1] show that it is not only the quality of arguments that matters but also the subjective affinity influenced by position currently held. If a sender holds opinion which falls within latitude of acceptance of a receiver, the latter will shift into direction advocated by the former. If it falls within latitude of rejection, a receiver will shift away from advocated position. Finally, it a message does not qualify to be accepted but neither is falls within latitude of rejection, a receiver will remain non-committed and no shift will take place [Jager and Amblard 2004: 295-296]. The main point is that people make judgments about contents of messages sent to them.

The Social Judgment Theory "proposes that persuaders must carefully consider the pre-existing attitudes an audience might hold about a topic before crafting a message. If you send a message that falls in a receiver's latitude of rejection, you will not be successful in your persuasive effort. Moreover, if you send a message that is clearly in a person's latitude of acceptance, you are not persuading that receiver, you are only reinforcing what she or he already believes. True persuasion can only occur, according to this theory, if the message you send is in an individual's latitude of non-commitment or at the edges of his/her latitude of acceptance" [Dainton and Zelley 2005: 108-109]. In political campaign, candidates don't try to persuade voters to accept extreme views but rather they want people to vote for them. Therefore, positions placed around the middle of a scale are usually accepted or, at least, "positively neutral" [O'Keefe 1990: 39].

These findings suggest that the intensity of campaign, assuming that many advocates hold extreme positions, can indeed produce effects contrary to those intended.

3.3. Agent-Based Modeling

The main idea of the model presented here is application of the Social Judgment Theory to assess dynamic changes of opinions and formation of opinion clusters within artificially created representation of voters and after multiple runs of simulation. Agent-based modeling has been used by number of researchers to assess dynamics of the Social Judgment Theory.

One of the most prominent works was conducted by Wander Jager and Frédéric Amblard [2004]. In their model, each individual agent has its own opinion ranging from -1 to 1, latitude of acceptance and latitude of rejection. During each time-step agents are paired on random basis and the exchange of opinions takes place. The strength of influence is controlled by parameter μ, which is set at 0.1:

$$if\ |x_i - x_j| < u_i,\ \text{then}\ dx_i = \mu(x_j - x_i);$$
$$if\ |x_i - x_j| > t_i,\ \text{then}\ dx_i = \mu(x_i - x_j);$$

where:

x_i, x_j – opinions of agent i and agent j;
u_i – threshold of acceptance of agent i;
t_i – threshold of rejection of agent i;
$t_i > u_i$.

Authors run simulation for 400 agents endowed with different opinions and sets of latitudes. The tests include manipulating latitudes in order to assess emerging clusters of opinions in spaces populated with particular types of agents.

H. F. Chau and colleagues [2013] build on Jager's and Amblard's model and introduce some interesting changes. They clarify that convergence parameter $\mu \in (0, 0.5]$ and they introduce divergence parameter $\lambda > 0$. Therefore, Jager's and Amblard's function of rejection is amended to the form:

$$if\ |x_i - x_j| > t_i,\ \text{then}\ dx_i = \lambda(x_i - x_j).$$

The introduction of divergence parameter intends to solve the problem of doubtful validity resulting from the fact that $T > U$ implies greater magnitude of opinion change in the latitude of rejection than in the latitude of acceptance. Authors set μ at 0.2 and λ at 0.05.

The results which authors of both articles reach and assumptions they make are used in this model to create desired profiles of voters and rules of communication between them.

4. THE MODEL

The main idea behind the model was to introduce different types of agents that would communicate on random basis and update their opinions on Scottish independence. Dynamics of interaction is based on the Social Judgment Theory. This framework makes it easy to introduce varying levels of intensity of campaign, corresponding to the chance of interaction among agents at every time-step. Former sections of this paper mentioned several assumptions to be tested by the model:

--the trend of support for independence in Scotland is constant and it is not likely to change upon September 2014 referendum;

--aggressive campaigns of either camp can "scare" moderate voters and make them either unwilling to attend referendum at all or confuse them about the choice they should make;

--appealing to swing voters and those undecided should be the main point of campaigns.

4.1. Agents

The table below shows support for independence retrieved from multiple funders opinion polls by Scottish Social Attitudes[1]. Polls included five questions with distinction over membership in the European Union and keeping current devolution mechanisms. Here opinions are merged, summarizing those in favor of Scottish independence and those supporting remaining the part of the Great Britain.

[1] www.whatscotlandthinks.org/.

Table 1. Support for Independence in Scotland, 1999-2012

Year	Independent	Part of the GB	Don't know
1999	28%	68%	5%
2000	30%	67%	3%
2001	27%	69%	4%
2002	29%	64%	6%
2003	25%	68%	6%
2004	32%	62%	5%
2005	34%	58%	8%
2006	30%	63%	7%
2007	24%	71%	5%
2009	27%	64%	7%
2010	24%	71%	5%
2011	32%	64%	5%
2012	24%	72%	5%

The table below shows the share of votes each party received in consecutive elections to the Scottish parliament. There are two important modifications introduced. First, parties are grouped together along attitudes towards independence. Therefore, Scottish National Party is presented separately while Liberal Democrats, Labour and Conservative parties are grouped together picturing overall electoral support for pro-union parties. Second, the percentage of votes received by each block is multiplied by the corresponding turnout, therefore presenting the actual share of population eligible to vote and supporting either political block.

Table 2. Support for Political Parties in Scottish Parliamentary Elections, 1999-2011

	1999	2003	2007	2011
SNP (*independence*)	29%	24%	33%	45%
Labour (*union*)	39%	35%	32%	32%
Conservative (*union*)	16%	17%	17%	14%
Liberal Democrats (*union*)	14%	15%	16%	8%
Pro-unionists total	69%	67%	65%	54%
Turnout	59%	49%	52%	50%
Actual SNP support	17%	12% *(min)*	17%	23%
Actual pro-unionists support	41%	33%	34%	27% *(min)*

Data presented in both tables constitutes the fundament for agents' characteristics and distribution. One of the assumptions retrieved from the former research was that supporters of either political block don't necessarily shift their attitudes towards independence along party lines, even though some influence is possible. This model assumes that people who always vote in

elections and always pick the same party block are also always supporting either independence or union. The model names these voters "Hard Core Pro-Independence" and "Hard Core Pro-Union" agents respectively. Comparison of elections results and polls data shows that the lowest actual percentage of the population voting for Scottish National Party equaled 12 while those choosing pro-unionists constituted 27 percent. Both values are lower than results obtained from opinion polls, where minimum support for independence reached 24 percent and for union - 58 percent. Therefore, the total share of "Hard Core Pro-Independence" and "Hard Core Pro-Union" agents is 12 and 27 percent respectively. Summing up, both types of agents constitute a "hard core" in the sense that they always vote for the same party block, they always take part in national elections and they always take the same stand towards independence. They are likely to express "aggressive" arguments when communicating with other agents.

Another fact revealed by opinion polls data is that the percentage of those who were undecided towards the issue of independence ranged from 3 to 8 percent. The model introduces "Undecided" type of agents who, according to the Social Judgment Theory, are highly uncommitted to either case.

Finally, the remaining part of population with subtracted both hard core and undecided types constitutes "Shifting Independence" agents who may express either pro or anti-union attitudes but who are eager to change their options when confronting other arguments. Both "Undecided" and "Shifting Independence" types of agents may or may not vote is national elections, depending on the turnout.

The table below shows a summary of agents' characteristics with reference to the Social Judgment Theory and its application adopted by Jager and Amblard [2004], with opinions ranging [-1, 1] and U and T values chosen from (0, 2].

Table 3. Summary of Agents' Characteristics

Agent type	X	U	T	Party affiliation	Always vote?
Hard Core Pro-Independence	(0.5, 1]	0.5	0.7	SNP	yes
Hard Core Pro-Union	[-1, -0.5)	0.5	0.7	Unionists	yes
Shifting Independence	[-1, 1]	1.2	1.4	Changing	no
Undecided	0	0.2	1.6	Changing	no

Agents' opinions range from 1 for those supporting independence to -1 for union's enthusiasts. 0 represents an ideal point of neutrality while [-0.5, 0) and (0, 0.5] picture "positive neutrality" towards the union and independence respectively. The implication of "positive neutrality" is rather closeness to 0 and to the contrasting option than willingness to vote in referendum, although these agents may decide not to vote because of insufficient information about the issue, as it was indicated in the former research. If agents with opinion within this range decide to vote, they will choose accordingly to their X value. By assumption, the only voters who will always take part in referendum are hard core type of agents. Agents with X = 0 will decide either not to vote or will return a blank or invalid ballot.

The closeness of U and T values in the case of hard core agents means that they are highly ego-involved on the issue of independence. Jager and Amblard [2004] pointed out that in population inhabited by these agents the contrast effect dominates. "Shifting Independence" type of agents has high acceptance latitude with moderate non-commitment and rejection. Here assimilation takes place. Those undecided are significantly resisting opinions of others and they are highly non-committed. According to the tests, they will contrast themselves from the extremes [Jager and Amblard 2004: 297-300].

Upon each simulation both "Shifting Independence" and "Undecided" types of agents are randomly assigned to either pro-independence or pro-union political block, if they fall within random selection depending on chosen popular support for either camp. As party affiliation may or may not have influence on supporting particular attitude towards independence, randomly chosen modification of X ϵ [-0.5, 0] or [0, 0.5] is assigned. Values of U and T are considered inherent to every agent and don't change throughout a simulation. Finally, the normalization function is applied to every agent in order for X not to exceed either -1 or 1.

4.2. Simulation

The model space was populated with 100 agents corresponding to 100 percent of Scottish electorate. Upon each simulation, 27 agents were assigned to "Hard Core Pro-Union" type, 12 to "Hard Core Pro-Independence", "Undecided" included between 3 to 8 agents and "Shifting Independence" type was constituted by the remaining number of agents. After setting X, U and T values, assigning party affiliation and applying normalization function to X, multiple simulations were run. Each agent had a chance of establishing connection with randomly chosen other agent, depending on a predefined probability of interaction per day. One simulation consisted of 1460 days, corresponding to 4 years of electoral cycle, after which party affiliation and distribution of "Shifting Independence" and "Undecided" was reset. The exchange of opinions took place within each connection between agents and according to Jager's and Amblard's [2004] function with Chau's et al. [2013] divergence parameter added for rejection procedure. Convergence parameter was set at 0.1 while divergence parameter was set at 0.05, borrowing from the logic introduced in abovementioned works. The results were summarized for 100 simulations and for emerging clusters of opinions within ranges [-1, -0.5), [-0.5, 0.5], (0.5, 1].

5. RESULTS AND V&V ACCOUNT

The verification account assumed testing the model with sensitivity analysis in order to determine its behavior under manipulated conditions. In addition, verification and validation steps were followed on every step of model's development to assure correspondence with underlying assumption retrieved from the Social Judgment Theory and the former research. In order to validate the model, simulations were run for historic data retrieved from opinion polls and elections results. Outcomes were checked against these conditions and tested with equivalency testing using 20% criterion [Rogers et al. 1993]. Finally, multiple simulations were run for the current data aiming to both explain the dynamics of opinion changes and to test assumptions stated earlier in the paper.

The 1999-2003 electoral period was characterized by highly unequal party distribution, where Scottish National Party gained actual 17 percent support of population and pro-union parties

enjoyed 41 percent of support. Opinion polls from 2003 revealed 25 percent support for independence, 68 percent support for the union and 6 percent of undecided responses. Under conditions of the lack of political campaign, namely with the chance of interaction among agents set on 50 percent every day, results of 100 simulations revealed the average mean of 24 percent of supporters of independence, 71 percent of opponents and 5 percent of undecided. Standard deviations reached 8, 8 and 2 respectively. While average means indicated satisfactory validation results, standard deviations suggested much dispersion.

Equivalency test for independence support revealed that 90% confidence interval (-1.54 to 4.32) is contained within the equivalency interval (-5 to 5) using the 20% criterion. The difference between observed and simulated support for independence was within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was not rejected with a 5% risk of a Type I error. For union support, equivalency test showed that 90% confidence interval (-5.87 to 0.01) was contained within the equivalency interval (-13.6 to 13.6) using the 20% criterion. The difference between observed and simulated support for independence was within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was not rejected with a 5% risk of a Type I error.

During 2003-2007 electoral cycle the support for Scottish National Party was set on 12 percent, while pro-unionists enjoyed 33 percent of support. In 2007 there was 24 percent of population expressing pro-independence views, while pro-unionists constituted 71 percent of responses. Those undecided reached 5 percent. Simulation results revealed the average mean of 23 percent of independence supporters, 71 percent of pro-unionists and 5 percent of undecided, with standard deviations of 8, 8 and 2 respectively.

Equivalency test for independence support revealed that 90% confidence interval (-2.46 to 3.58) was contained within the equivalency interval (-4.8 to 4.8) using the 20% criterion. The difference between observed and simulated support for independence was within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was not rejected with a 5% risk of a Type I error. For union support, equivalency test showed that 90% confidence interval (-3.29 to 2.77) was contained within the equivalency interval (-14.2 to 14.2) using the 20% criterion. The difference between observed and simulated support for independence was within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was not rejected with a 5% risk of a Type I error.

The 2007-2011 electoral period was characterized by 17 percent support for Scottish National Party, 34 percent of support for pro-unionists and the percent of independence supporters reaching 32 in 2011, 64 of pro-unionists and 5 of undecided. Simulation revealed average mean of 24 percent of supporters of independence, 70 percent of pro-unionists and 6 percent of undecided, with corresponding standard deviations of 9, 9 and 2. Here, the discrepancy between actual and simulated support for independence was evident and simulated outcomes correspond more to opinion polls from 2010, where the support for independence reached 24 percent and for union – 71 percent. The model was unable to assess the significant discrepancy between 2010's and 2011's 8 percent leap in support for independence.

Equivalency test for independence support revealed that 90% confidence interval (4.70 to 10.85) was not contained within the equivalency interval (-6.4 to 6.4) using the 20% criterion. The difference between observed and simulated support for independence was not within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was rejected with a 5% risk of a Type I error. For union support, equivalency test showed that 90% confidence interval (-9.13 to -2.95) was contained within the equivalency interval (-12.8 to 12.8) using the 20% criterion. The difference between observed and simulated support for independence was within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was not rejected with a 5% risk of a Type I error.

Additional equivalency test was run for 2010 results for pro-independence outcomes. It showed that 90% confidence interval (-3.296 to 2.86) was contained within the equivalency interval (-4.8 to 4.8) using the 20% criterion. Here, the difference between observed and simulated support for independence was within 20% of observed mean. The null hypothesis of no difference between observed and simulated means was not rejected with a 5% risk of a Type I error.

Table 4. Comparison of Historic Data and Simulation Results

Cycle	**1999-2003**	**2003-2007**	**2007-2011**	***2010***
SNP support	17%	12%	17%	17%
Unionists support	41%	33%	34%	34%
Historic pro-independence	25%	24%	32%	24%
Historic pro-union	68%	71%	64%	71%
Historic undecided	6%	5%	5%	5%
Simulated pro-independence	24%	23%	24%	24%
Simulated pro-union	71%	71%	70%	70%
Simulated undecided	5%	5%	6%	6%

Table 5. Equivalency Test for Difference between Observed Outcomes and Simulated Means Using 20% Criterion

Cycle	Ind. / Union	Lower confid. limit	Upper confid. limit	δ_1	δ_2	H_0 (no difference)
1999-2003	Ind.	-1.53	4.31	-5	5	Not rejected
1999-2003	Union	-5.87	0.01	-13.6	13.6	Not rejected
2003-2007	Ind.	-2.46	3.58	-4.8	4.8	Not rejected
2003-2007	Union	-3.29	2.77	-14.2	14.2	Not rejected
2007-2011	Ind.	4.7	10.86	-6.4	6.4	Rejected
2007-2011	Union	-9.13	-2.95	-12.8	12.8	Not rejected
2010	Ind.	-3.296	2.86	-4.8	4.8	Not rejected

Simulation results demonstrated that party affiliation has a little effect on attitudes towards independence. Even though agents were endowed with potentially positive change in their views on the issue after voting for either camp, the communication factor was responsible for ultimate shape of attitudes.

Since 2011, the distribution of support for Scottish National Party and pro-unionists has been 23 and 27 percent respectively. The table below shows average means of 100 simulations run for this distribution and assuming diverging chances of interaction. These changes aim to correspond to various levels of intensity of political campaign.

Table 6. Simulation Results for Diverging Intensity of Campaigns

Intensity	50%	70%	85%	100%
Simulated pro-independence	25.18%	25.33%	25.62%	25.47%
SD	9.62%	8.84%	8.66%	9.08%
Simulated pro-union	69.17%	68.97%	68.86%	69.22%
SD	9.52%	8.86%	8.88%	8.66%
Simulated undecided	5.65%	5.7%	5.52%	5.31%
SD	1.58%	1.61%	1.71%	1.82%

Simulated support for independence was slightly higher than under previous party distribution however there was no clear relation between increasing share of seats in parliament and emerging support for independence. What is more, the intensity of campaign under given party distribution had virtually no influence on the support for either case. It clearly suggests that no matter if campaign is aggressive or not, it can neither damage nor help in forcing either option.

In order to make the results more informative, the table below shows emerging average means of clusters of opinions under 2011 party distribution and assuming two extreme chances of interaction.

Table 7. Average Means of Opinion Clusters

Intensity	50%	100%
x > 0.5	0.9963	0.9962
SD	0.0045	0.0042
x < -0.5	-0.9862	-0.9867
SD	0.0094	0.0087
$-0.5 \leq x \leq 0.5$	0.0366	0.0298
SD	0.0623	0.05

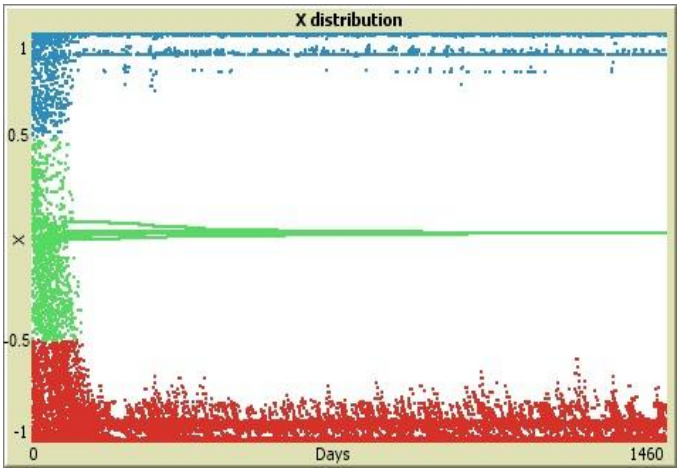


Figure 1. Clusters of X Values for Sample Simulation with Intensity of Campaign Set On 50%

Once again, the chance of interaction changed very little when it comes to distribution of opinions. The emergent phenomenon was three main clusters, with pro-independence and pro-union opinions grouped around extreme 1 and -1 points respectively. Those undecided remained very close to 0 option. Given that "Undecided" type is unlikely to exceed 8 percent of votes, the relative "positive neutrality" towards independence should matter little for political campaigners. It seems that regardless of what pro-union camp does, it will be able to grasp sufficient majority of voters supporting the union, especially if one assumes increased chance of voting in referendum when voters approach either of extreme options.

It was interesting to check how hypothetical distribution of parties changes the support for independence. Assuming that

Scottish National Party was able to grasp 41 percent of population's support and pro-union parties secured 17 percent, the situation mirrors the one from 1999, with the turnout set on 59 percent. The table below shows the results for two extreme values of intensity of campaign under these conditions.

Table 8. Simulation Results for Hypothetical Distribution of Parties

Intensity	**50%**	**100%**
Simulated pro-independence	28.08%	31.22%
SD	11.57%	11.43%
Simulated pro-union	66.27%	63.44%
SD	11.13%	11.23%
Simulated undecided	5.65%	5.34%
SD	1.67%	1.6%

The intensified campaign had some positive effect on the support for independence however the independence camp was still unable to grasp the sufficient number of votes, even within one standard deviation.

All these results suggest that the support for either option relies more on inherent distribution of hard core types of agents within the population. Regardless of the intensity of campaign, the prevalent loyal representatives of either camp are able to form clusters of agents shifting their attitudes easily. One last thing worth experimenting with was a different distribution of both "Hard Core Pro-Union" and "Hard Core Pro-Independence" types of agents. The table below shows results for electorate populated with 20 percent of hard core unionists and 19 percent of hard core pro-independence agents.

Table 9. Simulation Results for Hypothetical Distribution of Agents

Intensity	**70%**	**100%**
Simulated pro-independence	50.67%	52.67%
SD	15.76%	15.35%
Simulated pro-union	43.96%	41.72%
SD	15.81%	15.28%
Simulated undecided	5.37%	5.61%
SD	1.77%	1.69%

Here, the chances for either camp to force its option in a referendum were more equal, even though the high standard deviation points out a significant portion of uncertainty. In this particular environment, the political campaign could direct undecided voters which might constitute a critical margin required to win. It is important to notice that in this hypothetical distribution pro-independence hard core agents were still in relative minority to pro-unionists but the outcome was nevertheless the support for independence. It suggests that there is certain threshold situated close to equality of distribution of extreme opinions holders, where option of minority is favored. This finding supports theoretical assumptions of the Social Judgment Theory.

6. CONCLUSION AND FURTHER RESEARCH

It is fair to agree with opinion polls expert Nate Silver that the data on support for Scottish independence is fixed and definitive. Indeed, simulation results show that the main target of campaigns, namely those Scotsmen who shift their attitudes towards independence, tend to cluster around either of X extremes, while those undecided are both too few in number to change possible outcome of the referendum and too rigid in maintaining their inconclusive opinions. Those who shift are simply "caught in the orbit" of whatever option has more hard core representatives. Average means of $X \in (-0.5, -1]$ and $(0.5, 1]$ reveal that by the end of electoral cycle both clusters are well-established. On the Scottish National Party's side, pro-independence campaign yields poor results no matter how intense it is, while pro-unionists don't have to do much in order to convince population to reject independence. Only when the distribution of hard core type of agents approaches equality, the chances for independence gain significance. Since this distribution relies mostly on previous voting behavior, more general elections are probably needed to "produce" more devoted independence supporters. This however is beyond the scope of the model which assumes fixed values of U and T.

Results revealed by simulation allow to conclude about three assumptions stated earlier in the paper:

-- "the trend of support for independence in Scotland is constant and it is not likely to change upon September 2014 referendum" – the statement is confirmed under given conditions;

--"aggressive campaigns of either camp can 'scare' moderate voters and make them either unwilling to attend referendum at all or confuse them about the choice they should make" – the statement is rejected as two big clusters emerged while the cluster formed within $X \in [-0.5, 0.5]$ was small and insignificant;

--"appealing to swing voters and those undecided should be the main point of campaigns" – the statement is rejected under given conditions as without more equal distribution of hard core type of agents campaigns are of secondary importance.

Certain limitations of the model point out considerations to be paid by the readers of this paper and show further ways to improve its explanatory and prediction powers. First, U, T and X values are assigned arbitrarily with rationale based primarily on earlier work by Jager and Amblard [2004]. Future work should include data gathering, aiming to assess more validated X, U and T values, possibly grouped along various constituency lines, age or territorial distribution. Even though current model yields satisfactory results when it comes to average means of opinions under 20% equivalency criterion, more exact data might help in obtaining satisfactory results for equivalence test with criterion lowered to 10%.

Second, for simplicity reasons, the model assumes that intensity of campaign depends on the chance of interaction between agents each day. Further work should introduce additional variables controlling the intensity and scope of campaign.

References

Carrell, Severin. "Scottish Independence: The Essential Guide." *The Guardian*, April 23, 2013.

Carrell, Severin. "Alex Salmond: Vote for Scottish Independence Is Act of Self-Belief." *The Guardian*, October 18, 2013.

Carrell, Severin. "Opinion Polls Show Voters Can Be Steered to Say 'Yes'." The Guardian Scottish Independence Blog, September 4 2013, http://www.webcitation.org/query?url=http%3A%2F%2Fwww.theguardian.com%2Fpolitics%2Fscottish-independence-blog%2F2013%2Fsep%2F04%2Fscotland-independence-opinionpolls-panelbase&date=2013-11-09.

Carrell, Severin. "David Cameron Urges Scottish Tories to Be More Aggressive in Their Politics." *The Guardian*, March 23, 2012.

Chau, H. F., C. Y. Wong, F. K. Chow and Chi-Hang Fred Fung. "Social Judgment Theory Based Model on Opinion Formation, Polarization and Evolution." 1-8: University of Hong Kong, 2013.

Dainton, Marianne and Elaine D. Zelley. *Applying Communication Theory for Professional Life: A Practical Introduction*. Thousand Oaks, CA: SAGE Publications, 2005.

Denver, David, James Mitchell, Charles Pattie and Hugh Bochel. *Scotland Decides: The Devolution Issue and the Scottish Referendum*. Abington: Routledge, 2000.

Dinwoodie, Robbie. "Support for Independence Growing." *Herald Scotland*, June 8, 2011.

Higgins, Charlotte. "Scottish Independence Campaign Has Almost No Chance: Says Nate Silver." *The Guardian*, August 13, 2013.

Jager, Wander and Frédéric Amblard. "Uniformity, Bipolarization and Pluriformity Captured as Generic Stylized Behavior with an Agent-Based Simulation Model of Attitude Change." *Computational & Mathematical Organization Theory* 10 (2004): 295–303.

Laycock, Samantha. "Is Referendum Voting Distinctive? Evidence from Three Uk Cases." *Electoral Studies* 32 (2013): 236–52.

McCrone, David. "Scotland out the Union? The Rise and Rise of the Nationalist Agenda." *The Political Quarterly* 81 (2012): 69-76.

McLaughlin, Mark. "Scottish Independence: Undecided Voters at New High." *The Scotsman*, October 11, 2013.

Mina, Denise. "Does Scotland Want Independence?" *The New York Times*, October 2, 2013.

O'Keefe, Daniel J. *Persuasion: Theory and Research*. Newbury Park, CA: Sage, 1990.

Peterkin, Tom. "Scottish Independence Support at 25 Per Cent - Poll." *The Scotsman*, October 12, 2013.

Qvortrup, Matt. "The 'Neverendum'? A History of Referendums and Independence." *Political Insight*, September 2013.

Rogers, James L., Kenneth I. Howard and John T. Vessey. "Using Significance Tests to Evaluate Equivalence between Two Experimental Groups." *Psychological Bulletin* 113 (1993): 553-65.

Sherif, Muzafer and Carl I. Hovland. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*. New Haven: Yale University Press, 1961.

Taylor, Brian. "Scottish Devolution." *BBC*. http://www.webcitation.org/query?url=http%3A%2F%2Fwww.bbc.co.uk%2Fnews%2Fspecial%2Fpolitics97%2Fdevolution%2Fscotland%2Fbriefing%2Fscotbrief1.shtml&date=2013-11-09.

Wintour, Patrick. "Amid the Chaos, Scotland Takes Historic Step." *The Guardian*, May 4, 2007.

Wray, Ben. "Election Analysis: Why the Nats Dismantled Labour in Scotland." *Counterfire*, May 6, 2011, http://www.webcitation.org/query?url=http%3A%2F%2Fwww.counterfire.org%2Findex.php%2Farticles%2Fanalysis%2F12184&date=2013-11-09.

Biography

Jan Nalaskowski is PhD candidate at Graduate Program in International Studies, Old Dominion University in Norfolk, Virginia. He specializes in comparative studies as well as modeling and simulation. His main areas of interest include Europe, European Union, independence movements, foreign behavior of subnational and unrecognized entities, system dynamics and chaos theory. He completed his coursework at Nicolaus Copernicus University in Torun, Poland, Stockholm University and University of Oxford. He is a Research Fellow of Casimir Pulaski Foundation, member of International Studies Association and The Society for Modeling & Simulation International. Jan is a recipient of Fulbright Graduate Student Award.

Agent-Based Simulation Event Execution Architecture for Improved Performance and Scalability

Jesse Caldwell, Tyrell Gardner, Dr. Jim Leathrum
Department of Modeling, Simulation, and Visualization Engineering at Old Dominion University
[jcadl013, tgard011, jleathru]@odu.edu

Keywords: Agent, Simulation, Architecture, Agent-Based Simulation, Event List, Discrete Event Simulation

Abstract: This paper describes the conceptual design and performance results of a simulation architecture. The simulation architecture is designed for use in agent-based simulations where the agents exhibit some logical grouping or organization. This is demonstrated in this work through a spatial grouping forming a gird. The grid structure is further organized into a hierarchy, containing one or more levels. The hierarchy is then utilized to create a hierarchical event management structure. This type of event management structure allows the simulation architecture to scale to problems with more than 10^6 agents. This paper provides a high-level description of the simulation architecture and presents the performance results of the architecture in various scenarios.

I. Introduction

This paper introduces a simulation architecture that aims to provide a solution to problems that can be described with an agent-based model and exhibits some logical grouping. The simulation architecture was developed to support a discrete-event epidemiology simulation application. The application took what is traditionally a continuous model and discretized it as an agent-based model. This paper presents the conceptual design and performance of the simulation architecture. The paper is divided into five sections. Section II presents the problem the simulation architecture aims to solve. Section III outlines the conceptual design of the simulation architecture. Section IV presents the performance results of the simulation architecture. Section V is the conclusion.

II. Problem Description

Agent-based simulations are useful for solving problems that are "characterized by the presence of a number of autonomous entities [or agents] whose behaviors (actions and interactions) determine the evolution of the system." [1] Within this set of problems exists a subset in which the entities exhibit some logical grouping with one another, either spatially or non-spatially. Figure 1 shows an example of agents with these properties organized on a grid. This is the set of problems the presented simulation architecture aims to solve. The simulation architecture also aims to provide a framework that is scalable from problems with a small number of agents (about 10^3 agents) to a large number of agents (10^6 agents or greater).

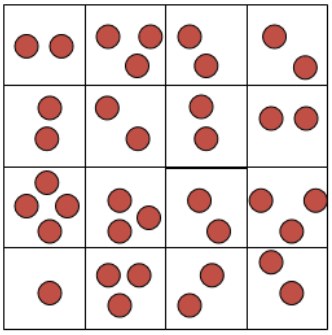


Figure 1. Sample grid of agents organized into logical groups based on proximity.

One of the primary results that can be abstracted from the logical organization of agents is the ability to create a hierarchy of the groups of agents. Assuming the groups of agents are organized into a grid, like in Figure 1, a clear hierarchy can be formed from separating the grid into four quadrants. This creates a layer of four groups that contain a set of four groups. If the groups are abstracted again into one group, the gird now becomes a tree as seen in Figure 2.

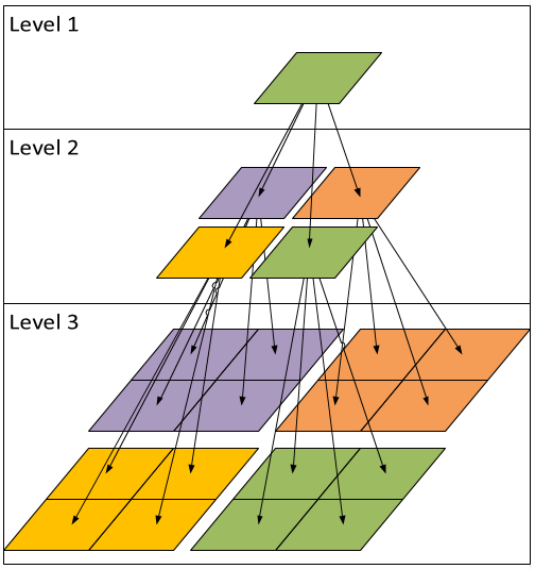


Figure 2. Three-level tree abstracted from the grid in Figure 1.

The agent-based simulations of interest operate using discrete-event simulation architecture. The agents within the simulation all perform actions at discrete points in time that result in a change in state of the agent or the system. One of the key aspects of agent-based simulation is the agent's interaction. Without the interaction with the environment and other agents, the global system dynamics would not emerge. This is because the global system dynamics are derived from the local behaviors and interactions of agents within the system [1].

III. Simulation Architecture Design

The simulation architecture was designed to improve the performance and scalability of agent-based simulations that have an inherent grouping. The simulation takes advantage of the agent's proximity and grouping to improve event management.

There are multiple strategies used to manage events in a discrete-event simulation. All event management strategies aim to address issues with event insertion, execution, and removal. Some strategies aim to limit the event execution process to a constant time operation, while others aim to reduce the event insertion process below O(n). Some of these strategies include a linear queue, heap, and calendar queue [2].

The agents in the simulation have behavior models which determine the actions the agents will take. Because of this, there are a limited number of actions that a given agent can take, where each action corresponds to an event in the simulation. Each agent in the simulation will keep a linear queue of the events that it will perform, which will tend to be small in size. Then agents only schedule their next event with this system.

By having each agent maintain its own event list, we can take advantage of the natural grouping of the agents. A group which holds agents, maintains an event list of the next executable event for each agent, thus a maximum of one event for each agent. Since the groups of agents are organized in a grid structure, we can abstract the grid into a hierarchical structure of groups (Figure 2), forming a tree. Each group in the tree maintains an event list that contains the event with the smallest execution time of each of its child groups. These groups' event lists are limited to having a maximum number of events equal to the number of children.

The insertion and removal of events in this hierarchy require special algorithms to utilize the hierarchy of the simulation. Figure 3 shows the algorithm for the event removal process.

```
void EventRemoval(String eventID)
{
    for(each event in the EventList)
    {
        if(current eventID == eventID)
        {
            if(no previous event)
            {
                head = next event

                parentGroup.EventRemoval(eventID);
            }
            else
            {
                connect previous event to next event
            }

            delete event
        }
    }
}
```

Figure 3. Event removal algorithm.

The algorithm is recursive to allow the tree to be altered if an event is removed from the top of the agent's event list. This algorithm is necessary to complete the event insertion process, because when an event is added at the top of the agent's event list

the event that was replaced must be removed from the agent's group event list and all parent group event lists the event was located. Figure 4 shows the algorithm for the event insertion process.

```
void EventInsertion(Event newEvent)
{
    if(newEvent execution time < execution time of the
                first event in the EventList)
    {
        insert newEvent as the first event in the EventList

        parentGroup.EventRemoval(id of next event);
        parentGroup.EventInsertion(newEvent);
    }
    else
    {
        iterator = first event of the EventList

        while(the next event is not NULL)
        {
            if(newEvent execution time <
                    next iterator event execution time)
            {
                inerst new event at iterator location
            }
            else
            {
                iterator = next event
            }
        }
    }
}
```

Figure 4. Event insertion algorithm

By creating this structure, the simulation is able to insert events, execute events, and remove events quickly. Due to the simplicity of the event lists, they are implemented by simple linear linked lists. The event insertion process within the agent is O(k) within the agent where k is the number of actions an agent can perform. Since inserting an event can require a modification of events in the tree as well, the event insertion process is O(log n) within the tree where n is the number of levels in the tree. The event execution process is O(log n). The event removal process is also O(k) with an O(log n) operation to modify the tree, since the events are local to the agents. This architecture's performance is limited by the number of agents in the groups that are a part of the grid. If the agents are evenly distributed about the grid and the number of agents in each group is small ($< 10^3$), then the architecture will perform well. However, if the agents are not distributed evenly and the number of agents in a group is large, then the architecture will suffer in performance.

The architecture also provides a benefit in scalability. The hierarchical structure offers the ability to add or remove levels in the hierarchy to best suit the situation being simulated. Another benefit of this architecture is the ability to set the number of children groups attached to each level of the tree. These two elements make the simulation scalable to very large problems (greater than 10^6 agents) without sacrificing performance.

IV. Architecture Performance

This section describes the application that was developed using the architecture to test the architectures performance. Then the simulation architecture's performance and a performance analysis are presented.

A. Application Description

The model used in this study is an agent-based model of infectious disease spread that is event driven rather than continuous time step driven based on the Kolmogorov forward equation [3]. Agents in the simulation represent individuals that have two state variables, disease state and position. The set of disease states is taken from the Susceptible-Exposed-Infected-Recovered (SEIR) infectious disease model shown in Figure 5 [4]. An agent can exist in one of these four disease states. The position state variable is represented as an ordered pair that can be plotted on a Cartesian plane. This plane can be viewed as an abstract grid with each cell representing a grouping of agents. The groups, otherwise referred to as grid cells, create an environment for the agents to interact and spread the disease.

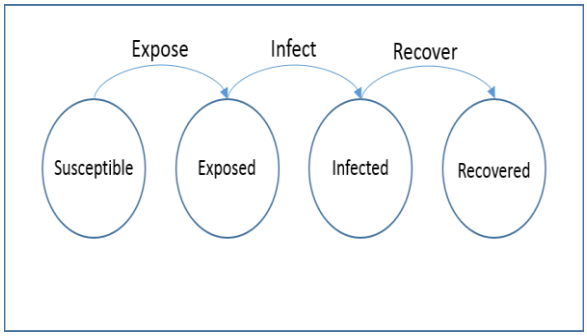


Figure 5. State Machine of Infectious Disease.

Examining Figure 3 reveals three events in the SEIR model, and these events are labeled *Expose*, *Infect*, and *Recover*. When one of these events occur the disease state variable of the agent is updated accordingly. The only other event that occurs in the model is *Move*. This event changes the position state

variable of the agent. In order to determine if certain events occur, a unique probability labeled the contact probability is used. This probability is used to determine if an infected agent will come in contact with a susceptible agent that is in the same grid cell. An infected agent is contagious and can expose susceptible agents to the disease, but in order for this to occur the agents must be in close proximity. If an infected agent moves to a neighboring cell, then the number of susceptible agents that are exposed to the virus is computed using the contact probability. The same thing happens in a reverse manner when a susceptible agent moves to a neighboring cell. The susceptible agent will signal all of the infected agents in the cell to determine if exposure takes place. As agents move between grid cells, the infectious disease will spread based on the infectious disease being modeled and the contact probability that is used.

B. Computational Performance

The application described above was implemented using the simulation architecture described in section III. For the performance tests, the simulation was run for 750 hours (simulation time), with two child groups per level in the hierarchy, a contact probability of 0.05%, and started with 5% of all agents infected. The two parameters that were controlled in the tests were the number of agents in the simulation and the number of levels in the tree, which implicitly changes the number of groups in the lowest level of the grid to 2^{n-1}, where n is the number of levels in the tree. The specifications for the computer used to gather performance results of the simulation were as follows: i5-2500k quad core processor, 8 GB of RAM, and Windows 7 Professional. The performance of the simulation was determined by the execution time of the simulation, while varying the number of agents and levels in the tree.

The performance measures were found by running the simulation with 10^N agents (N = 2,3,4,5,6) and M levels in the hierarchy (M = 1,2,3,…,8). Figure 6 shows a logarithmic plot of the results by the number of levels in the hierarchy. For each set of agents, a specific value of M demonstrates the best performance. This was an anticipated result due to the desire for the simulation architecture to be scalable to problems varying in size.

By examining the figure there are some

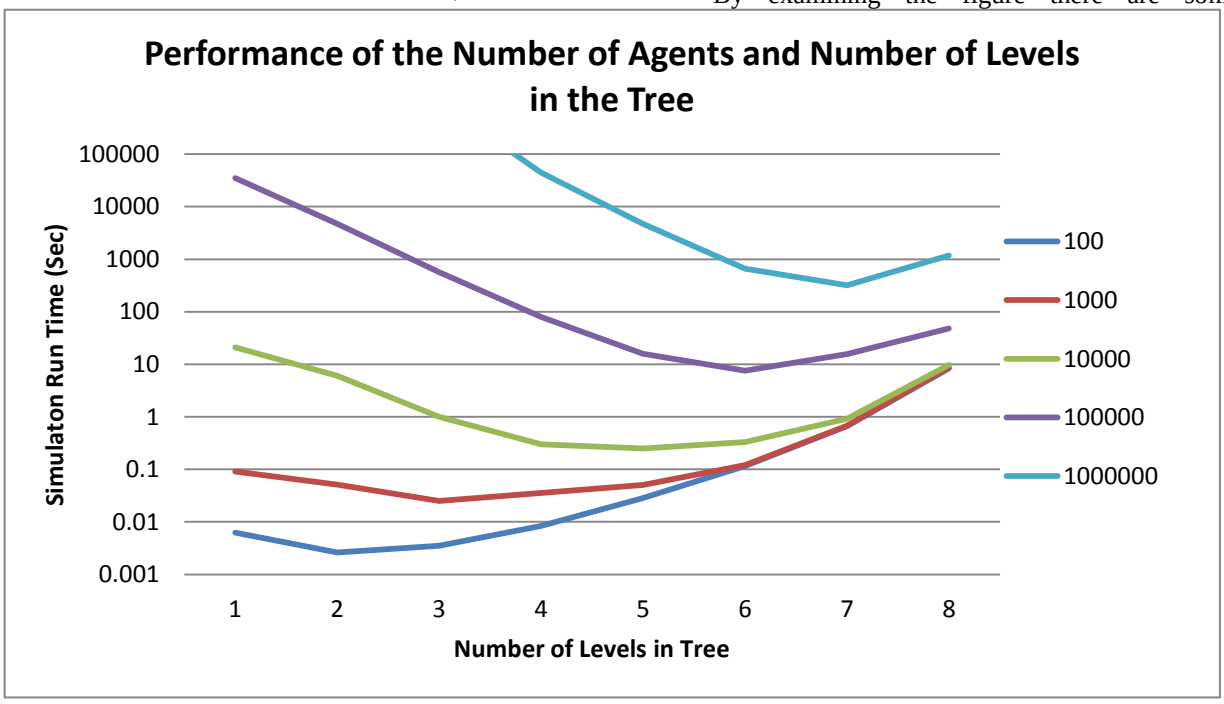


Figure 6. Graph of Simulation Performance

interesting behaviors which are exhibited. The first occurs when there are 10^4 agents or less. As the number of levels increase in the tree, they all converge to about the same performance time. This is expected since as the number of levels increase so does the number of groups in the grid (by 2^{n-1} for our test). As the number of groups increase, the agents become more evenly distributed about the grid to the point where there are more grid spaces than agents in the simulation. The performance of the simulation then suffers as agents become more thinly distributed about the grid, resulting in a grid computation dominating the agent computation, an undesirable effect.

The second behavior exhibits the diminishing returns on performance as the number of levels increases. For each set of agents simulated, an "optimal" number of levels in the tree exhibit the best result and each level added after that point has a slower performance time. The final behavior of interest of the simulation architecture is the capability to simulate 10^6 agents in approximately 10 minutes. This is significant because it would take longer than a week (calculated from extrapolating data via trend lines) to simulate the same number of agents with just one level in the tree. This is a significant speed up in the simulation performance, just by reducing the computational complexity from O(n) to O(log n). These results show the simulation architectures performance and scalability.

V. Conclusion

The simulation architecture described provides a framework for developing solutions to problems that can be defined as an agent-based model and exhibits some logical grouping or organization. The architecture not only performs quickly, but it is scalable to problems with large numbers of agents.

The current work on this simulation architecture focuses on agents organized in a grid. This is just one approach to take advantage of the natural organization of the agents, which focuses on a spatial organization. Other approaches could be taken that are either spatial or non-spatial; so long as the agents exhibit a natural organization this approach is valid and can be utilized.

There is still work that can be done to improve the performance and ease of use of the simulation architecture. The algorithms used to in the simulation architecture could continue to be optimized and refined. To make the simulation architecture easier to use, easier ability to customize and set the different system variables could be implemented. The team has been extremely pleased that the proposed simulation architecture works as well as it does.

Acknowledgements

The team would like to thank Dr. Jim Leathrum for his continued support, guidance, and expertise during the development of this project. Without him this project would not have been possible.

References

[1] Stefania, B., Manzoni, S., & Vizzari, G. (2009, 10 31). *Agent Based Modeling and Simulation: An Informatics Perspective*. Retrieved from JASSS: http://jasss.soc.surrey.ac.uk/12/4/4.html

[2] R. Brown, "Calander Queues: A Fast O(1) Priority Queue Implementaion for the Simulation Event Set Problem", *Communications of the ACM*, vol.31, no. 10, pp. 1220-1227, 1988

[3]Keeling, M., & Ross, J. V. (2008). On methods for studying stochastic disease dynmaics. *Journal of the Royal Society*, vol. 5, pp. 171-181.

[4] L. Perez and S. Dragicevic, "An agent-based approach for modeling dynamics of contagious disease spread." *International Journal of Health Geographic*, vol. 8, pp. 1-17, Aug. 2009.

Biography

Jesse Caldwell is a senior in the Modeling and Simulation Engineering program and Old Dominion University. His interests include serious gaming, formal methods, simulation software design, and simulation applications. Upon graduation in May 2014, he will be pursuing a job within the industry.

Tyrell Gardner is a Masters student in the Modeling and Simulation program. He is interested in medical simulations. Upon completion of his thesis in August 2014, he will be attending medical school at EVMS.

Dr. Jim Leathrum is an associate professor at Old Dominion University in the Department of Modeling, Simulation, and Visualization Engineering.

Modeling Effectiveness of Tick Control by a Species that Exhibits Predator-prey Role Reversal

Old Dominion University, Biological Sciences
Alexis White[1], Robyn Nadolny[1], Carrie Eaton[2], Holly Gaff[1]
[1]Old Dominion University, [2]Unity College
awhit130@odu.edu

Keywords: agent-based model, individual-based model, NetLogo, Ixodid ticks, Guineafowl, biological control

Abstract

Lyme disease, caused by the bacteria *Borrelia burgdorferi*, is the most commonly reported vector-borne disease within the United States. Blacklegged ticks (*Ixodes scapularis)* are the primary vectors of *B. burgdorferi* infections in humans. Because ticks have a complex life history and feed on several animal hosts throughout their lifetimes, it has historically been difficult to identify a method for effective control. Within the past 20 years, the use of domesticated guineafowl as a control on tick populations has increased in popularity. Guineafowl are known for their consumption of ticks and therefore many landowners assume that they provide protection from Lyme disease. However, there is very little evidence to support this claim; in fact, guineafowl have also been found to be a host for juvenile life stages of ticks. Mathematically, the relationship between guineafowl and ticks can be referred to as predator-prey role reversal, with guineafowl acting as both host and predator to ticks. To examine the biological control ability of guineafowl on tick populations, an agent-based model was developed. Initial results of the model support that guineafowl have only limited control over tick populations.

1. INTRODUCTION

Lyme disease, caused by the bacteria *Borrelia burgdorferi*, is endemic throughout the Northern Hemisphere (Higgins 2004). Within the United States, Lyme disease is the most commonly reported vector-borne disease and is most prevalent in the Northeast and Midwest areas of the country (Eisen et al. 2012). Cases of Lyme disease continue to grow annually. The need for both prevention and management is evident, but it is not an easy task (Gatewood et al. 2009; Ostfeld et al. 2006). Due to the complex life history of the tick species and their dependency on host interactions, an effective control option is difficult to find.

While Ixodid tick life histories vary from species to species, most consist of four specific life-stages: egg, larva, nymph, and adult. To progress from one stage to the next, ticks need a complete bloodmeal from a host (Sonenshine 1991). Host species for many larval and nymphal life-stages include birds, rodents, and reptiles (Tanner et al. 2010; Tijsse-Klasen et al. 2010; Eisen et al. 2004; Casher et al. 2002; Durden et al. 2002; Smallridge & Bull 1999; Apperson et al. 1993; Manweiler et al. 1990). The adult ticks tend to prefer large-bodied hosts such as white-tailed deer (*Odocoileus virginianus*) (Sonenshine 1991).

In an attempt to control the rise of Lyme disease and other zoonotic diseases vectored by ticks, such as Rocky Mountain spotted fever (*Rickettsia rickettsii*) or ehrlichiosis (*Ehrlichia chaffeensis*), researchers typically focus on reducing the abundance of ticks within an area (Ostfeld et al. 2006). In particular, there are two major forms of controlling tick populations: chemical control and biological control.

Chemical control techniques have focused on insecticides either administered to areas of the environment where a tick would seek for a host, or administering them on the host to not allow attachment and feeding of the tick. Addition of chemicals such as insecticides is not proven to be historically effective over long periods of time and is not ideal for the health of the environmental system. Therefore, direct biological control via predation is the preferred approach to reducing a species (Ostfeld et al. 2006).

Biological control methodology can be approached either from the top of the trophic structure of the bottom. Starting at the bottom of the trophic system for ticks involves manipulation of host populations. Modeling has shown reduction of host populations can be effective if the focus of management lies on hosts for the juvenile species such as larvae and nymphs (Buskirk & Ostfeld 1995). There are a multitude of biological control agents known to limit tick populations including bacteria, fungi, spiders, ants, beetles, rodents, and birds (Samish & Rehacek 1999). Helmeted guineafowl (*Numida meleagris*) are known for use as a domesticated control species on ticks (Fielden et al. 2008; Duffy & Brinkley 1992). Keeping these birds has increased in popularity over the past 20 years because of their presumed consumption of ticks and prevention of Lyme disease within backyards (Duffy & Brinkley 1992). However, there is very little evidence to support this claim;

in fact helmeted guineafowl may actually be an ideal host for juvenile life stages. A study conducted in South Africa collected data on a variety of wild bird species. A single guineahen was found with over 300 juvenile ticks attached and was the most heavily infested species of bird (Niekerk et al. 2006).

Ecologically, guineafowl do not become consumed by the ticks as ecological prey would, but mathematically the energetic exchange from host to parasite and parasite to host, allows for the term predator-prey role reversal to be utilized in this situation. Guineafowl act as both a host and a predator to ticks, which complicates their true ability to control tick populations simply based on number of ticks consumed. Other examples of role reversal between predator and prey are rare, but have been observed amongst other organisms such as ground beetles (Carabidae) and amphibians, or wood frogs (*Rana sylvatica*) and the spotted salamander (*Ambystoma maculatum*) (Wizen & Gasith 2011; Petranka et al. 1998).

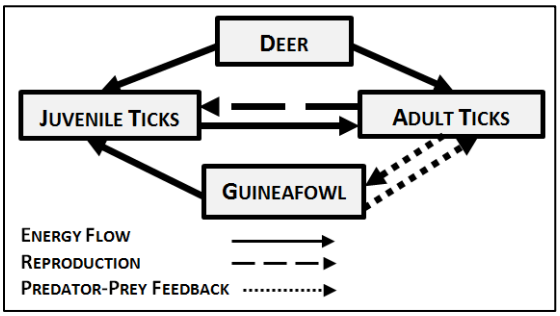


Figure 1. Diagram of predator-prey role reversal feedback with guineafowl and ticks.

One way to examine the interactions between ticks and guineafowl populations is through mathematical modeling. Recent tick studies have not included the role of predators in reducing tick populations, and no mathematical model of tick ecology has included predation as a significant source of mortality (Johnson et al. 2010; Preston & Johnson 2010; Mwangi et al. 1991). Within this study, predator-prey role reversal is modeled to test if guineafowl can control tick populations. Utilizing agent-based modeling within the software NetLogo, various population dynamics and interactions can be altered (Wilensky 1999). By using results from other studies and known factors about tick life cycles as parameters, our model can reveal the ability of guineafowl to control tick populations through predation. Alternatively, we investigate the inability of guineafowl to control, or even amplify the overall population size of ticks with the introduction of a new food source for juvenile ticks.

2. METHODS

2.1 The model.

The model description follows the ODD (Overview, Design concepts and Details) protocol for describing individual- and agent-based models developed by Grimm et al. (2006; 2010) and consists of seven elements. The first three elements provide an overview, the fourth element explains general concepts underlying the model's design, and the remaining three elements provide details.

2.1.1 Purpose.

The main purpose of this initial model is to explore population dynamics between tick, host, and predator. The scenario results will be compared to discuss whether a population which acts as host can also be an effective predator for biological control. The first scenario is used to determine the host population needed to sustain a tick population in a system without guineafowl. Using this value, guineafowl are introduced as a host to explore the impact of an additional host species on the tick population in the system. The final scenario is used to analyze the underlying purpose of the model, thus, guineafowl are introduced as both host and predator to determine their effectiveness as a biological control given varying capture efficiency as a predator.

2.1.2 State variable and scales.

In the agent-based model there are three types of agents: ticks, guineafowl, and deer. Each agent has given traits and is followed over successive time intervals. This initial model parameter values based loosely on the population dynamics of the blacklegged tick (*Ixodes scapularis)* and the white-tailed deer, (*Odocoileus virginianus)*. The environment is set up at 25x25 patches of equal quality with wrapping boundaries.

Each tick agent has a unique identification number, life stage (adult or juvenile), and amount of energy. Adult ticks gain more energy from a deer bloodmeal than from a guineafowl to reflect host preference ecology. Juvenile ticks however receive the same bloodmeal energy gain from both hosts. Guineafowl agents also have a unique identification number, a count of ticks they have consumed, as well as a count of ticks that have fed upon them. They are a constant population as this system is currently set up to be a domestic population and it is assumed owners would replace any lost fowl. Lastly, deer agents have a unique identification number and a count of ticks that have fed on them. The deer population is also constant and encompasses all hosts of tick species outside of the guineafowl

2.1.3 Design concepts.

The agents in the simulation interact as the ticks find hosts to use for blood meals, and the guineafowl consume ticks as a predator. The ticks can only sense hosts within their patch, and the probability of attaching and finding a host is factored into a background mortality parameter. All processes are stochastic for all runs, and there are no fitness

differences between agents. Host movement between patches is a random process for simplicity.

2.1.4 Input.

Each simulation is initiated in a uniform 25x25 patch grid with an initial 100 juvenile ticks (larvae and nymphs) and 50 adults randomly spread across the grid. All other parameters are given in Table 1.

2.2 Simulation experiments.

Three experiments were conducted for this initial analysis. The first experiment ran the basic model 100 times using a varying sized deer host population. The number of guineafowl agents was set to zero. The results were averaged to identify population trends and to establish the minimum number of deer agents needed to support a stable tick population. Counts of number of ticks in each life stage were recorded for each time step, for a total of 2000 time steps to avoid stochastic anomalies. Tick population persistence was measured by a tick count being greater than 0 at the 2000 time step. Once the population persisted for 90% of the runs, this number of deer was utilized for the second experiment. The second experiment ran the model 100 times using varying guineafowl population as hosts. This revealed the dynamics of guineafowl as an additional host in the system on the tick population. The results were averaged and plotted. This experiment was also utilized as a contrast for the final scenario. The third experiment ran the model 100 times once again, with limited variation of the guineafowl population, but with varying efficiency of tick capture as a predator species. The results were averaged and plotted.

The model was coded using the programmable modeling environment, NetLogo. This software was authored by Uri Wilensky in 1999 and is freely available (http://ccl.northwestern.edu/netlogo/).

3. RESULTS

The average results of the first experiment are shown in Figure 2, and Figure 3 shows the percentage of persisting tick population. The average number of ticks, with the plotted standard deviation of the means reveals persistence of tick populations occurs around 60 deer agents, therefore 60 deer were utilized as an initial condition in the second experiment.

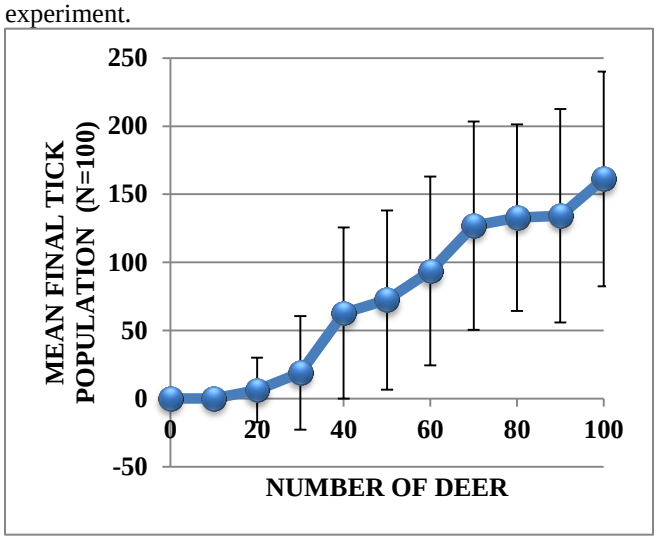


Figure 2. Results of experiment one, showing the mean of the final count of ticks given a varying deer population parameter. Standard error bars are used to show the variation around the means.

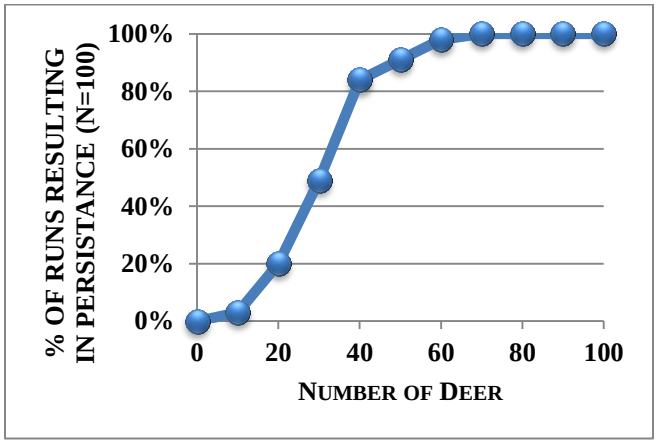


Figure 3. Graph of the first experiment results for the percentage of runs resulting in persistent tick populations (defined as a nonzero tick population after 2000 time steps).

The second experiment shows that there is no significant difference in the average population size of ticks given varying numbers of guineafowl in the system (Figure 4). Therefore the number of guineafowl in the third scenario was not a crucial number.

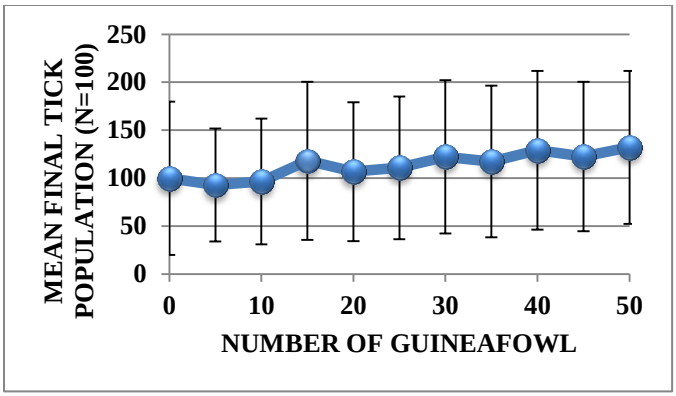


Figure 4. Mean final count of ticks given a varying guineafowl population parameter. Standard error bars are used to show the variation around the means.

The third experiment further supported that the number of guineafowl is not as important in determining the average number of ticks as the efficiency of predation (Figure 5). Also, given 100% predation efficiency tick populations continue to persist. There is a significant reduction in tick numbers, but no elimination of the tick population even under extreme conditions (Figure 6).

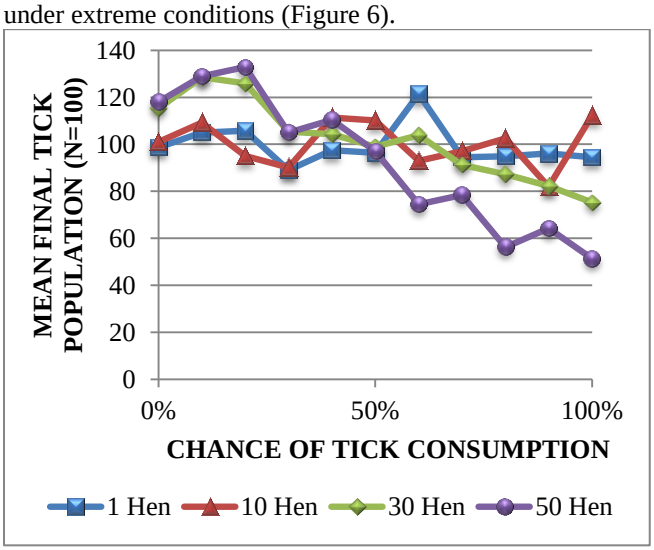


Figure 5. Mean final count of ticks given a varying guineafowl population and varying consumption parameter.

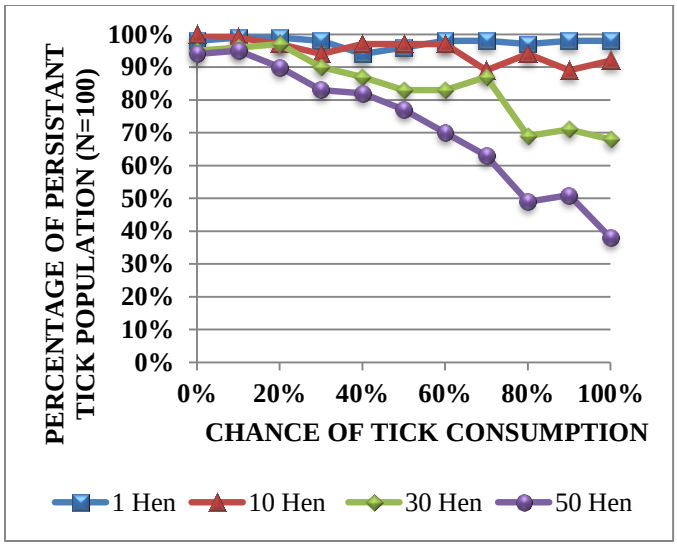


Figure 6. Percentage of runs resulting in a persistent ticks population with a varying guineafowl population and varying chance of tick consumption.

4. DISCUSSION

The simulation results for this model suggest that guineafowl do exhibit some control over tick populations, which supports findings from other studies (Duffy & Brinkley 1992). However guineafowl do not exhibit strong control over tick populations, and cannot eliminate ticks. The model is currently theoretical, with parameters based on model results and a literature review. The model would benefit from specific field data. There are many unanswered questions within the literature, such as the feeding preferences of guineafowl and if consuming ticks of different life stages has different effects on the tick population. There is also some question regarding the number of ticks that feed on guineafowl in domestic situations. Therefore it would be interesting to parameterize this model with field and lab data.

Furthermore, this simple model is based on a domestic habitat such as a homeowner's backyard. Guineafowl are only one species which acts as natural predators of ticks. There are also other species such as fence lizards or oxpeckers which are known for being both a predator and host of tick species (Casher et al. 2002; Samish & Rehacek 1999). Modifying the model to fit a different system, or more natural predator within a natural habitat would greater expand the knowledge gathered in this model.

Finally, ticks are known for the spread of the causative agents for many diseases. With the increase in global temperatures there has also been a noted increase in disease prevalence and tick population ecology is changing. There is a possibility that predator-hosts such as guineafowl may be able to control or reduce disease prevalence within an area even if they cannot eliminate tick populations. But this is all

based on the assumption that guineafowl are effective biological control agents for tick populations.

It may not be biologically realistic to assume that guineafowl can control or eliminate tick populations, given that guineafowl are not ticks specialists and ticks have incredible reproductive power and can feed on a diversity of hosts (Samish & Rehacek 1999). A fully engorged blacklegged tick (*I. scapularis*) can produce thousands of offspring (Sonenshine, 1991). If a guineafowl neglects to eat a single engorged female tick, thousands are reintroduced to the system and may even utilize the guineafowl as a host. Therefore, this example of a predator-prey feedback loop needs to be further studied, and further modeled. Once data collection and modeling are combined, informed management solutions can be applied to real systems, using efficient biological control to reduce tick populations.

5. ACKNOWLEDGEMENTS

A special thanks to all members of the ODU Tick Research Team for their support throughout this project. I would like to also thank my undergraduate thesis committee and university, Unity College, as this is a continuation and my thesis work.

6. LITERATURE CITED

Apperson, CS, JF Levine, TL Evans, A Braswell, J Heller. 1993. Relative utilization of reptiles and rodents as hosts by immature *Ixodes scapularis* (Acari: Ixodidae) in the coastal plain of North Carolina, USA. Experimental and Applied Acarology. 17:719-731.

Casher, L, R Lane, R Barrett, L Eisen. 2002. Relative importance of lizards and mammals as hosts for Ixodid ticks in Northern California. Experimental and Applied Acarology. 26:127-143.

Duffy, D, CR Brinkley. 1992. The effectiveness of helmeted guineafowl in the control of the deer tick, the vector of Lyme disease. Wilson Bulletin 104:342-345.

Durden, LA, JH Oliver, Jr., CW Banks, GN Vogel. 2002. Parasitism of lizards by immature stages of the blacklegged tick, *Ixodes scapularis* (Acari, Ixodidae). Experimental and Applied Acarology. 26:257-266.

Eisen, L, RJ Eisen, RS Lane. 2004. The roles of birds, lizards and rodents as hosts for the western black-legged tick *Ixodes pacificus*. Journal of Vector Ecology. 29:295-308.

Eisen RJ, J Piesman, E Zielinski-Gutierrez, L Eisen. 2012. What do we need to know about disease ecology to prevent Lyme disease in the northeastern United States? Journal Medical Entomology. 49.

Fielden, LJ, Y Rechav, NR Bryson. 2008. Acquired immunity to larvae of *Amblyomma marmoreum* and Ahebraeum by tortoises, guinea-pigs and guinea-fowl.Medical and Veterinary Entomology. 6: 251-254.

Gatewood, AG, KA Liebman, G Vourc'h, J Bunikis, SA Hamer, et al. 2009. Climate and Tick Seasonality Are Predictors of Borrelia burgdorferi Genotype Distribution. Applied and Environmental Microbiology. 75 (8): 2476-2483.

Grimm, V, U Berger, F Bastiansen, S Eliassen, V Ginot et al. 2006. A standard protocol for describing individual-based and agent-based models. Ecological Modelling. 198:115-126.

Grimm, V, U Berger, D DeAngelis, J Polhill, J Giske et al. 2010. The ODD protocol: a review and first update. Ecological Modelling. 221: 2760-2768.

Higgins, R. 2004. Emerging or re-emerging bacterial zoonotic diseases: bartonellosis, leptospirosis, Lyme borreliosis, plague. Rev Sci Tech. 23(2): 569–81.

Johnson, TJ, A Dobson, KD Lafferty, DJ Marcogliese, J Memmott, et al. 2010. When parasites become prey: ecological and epidemiological significance of eating parasites. Trends in Ecology and Evolution. 25: 362-371.

Manweiler, SA, RS Lane, WM Block, ML Morrison. 1990. Survey of Birds and Lizards for Ixodid Ticks (Acari) and Spirochetal Infection in Northern California Journal of Medical Entomology. 27:1011-1015.

Mwangi, EN, RM Newson, GP Kaaya. 1991. Predation of free living engorged female *Rhipicephalus appendiculatus*. Experimental and Applied Acarology. 12:153-162.

Ostfeld, RS, A Price, VL Hornbostel, MA Benjamin, F Keesing. 2006. Controlling ticks and tick-borne zoonosis with biological and chemical agents. Bioscience. 56: 383-394.

Petranka, J.W., A.W. Rushlow, M.E. Hopey. 1998. Predation by tadpoles of *Rana sylvatica* on embryos of *Ambystoma maculatum*: Implications of ecological role reversals by *Rana* (predator) and *Ambystoma* (prey). Herpetologica. 54(1): 1-13.

Preston, D, P Johnson. 2010. Ecological consequences of parasitism. Nature Education Knowledge. 1:39.

Samish, M, J Rehacek. 1999. Pathogens and predators of ticks and their potential in biological control. Annual Review of Entomology. 44:159-82.

Smallridge, CJ, CM Bull. 1999. Transmission of the blood parasite *Hemolivia mariae* between its lizard and tick hosts. Parasitology Research. 85:858-863.

Sonenshine, DE. 1991. Biology of Ticks, Vol. 1. Oxford University Press, New York.

Tanner, CL, FK Ammer, RE Barry, EY Stromdahl. 2010. Tick burdens on *Peromyscus leocopus* Rafinesque and infection of ticks by *Borrelia* spp. in Virginia. Southeastern Naturalist. 9:529-546.

Tijsse-Klasen, E, M Fonville, JH Reimerink, A Spitzen-van der Sluijs, H Sprong. 2010. Role of sand lizards in the ecology of Lyme and other tick-borne diseases in the Netherlands. Parasites and Vectors. 3:42.

Wilensky, U. 1999. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

Wizen, G., A. Gasith. 2011. An unprecedented role reversal: ground beetle larvae (Coleoptera: Carabidae) lure amphibians and prey upon them. PLoS one. 6(9).

Popularity or Proclivity? Revisiting Agent Heterogeneity in Network Formation

Xiaotian Wang and Andrew Collins

***Abstract*—In this study, the authors intend to apply agent-based modeling (ABM) approach to reassess the Barabasi-Albert model (BA model), the classical algorithm used to describe the emergent mechanism of scale-free network. The author argues that BA model as well as its variants rarely take agent heterogeneity into the analysis of network formation. In social networks, however, people's decision to connect is strongly affected by the extent of similarity. The author proposes that in forming social networks, agents are constantly balancing between instrumental and intrinsic preferences. Based on agent-based modeling, the author finds that heterogeneous attachment helps explain the deviations from BA model.**

I. Introduction

Social network analysis (SNA) has an especially long tradition in various disciplines of social science [1, 2, 3, 4]. First emerged in the 1930s, early SNA has focused primarily on the characteristics of individuals, assuming agent heterogeneity as the key factor in shaping the formation of different types of social networks [5, 6, 7]. In recent decades, the scholarly interest in SNA is reenergized by the proliferation of the information and communications technologies (ICTs) like the Internet and the mobile phone. A burgeoning literature has been devoted to exploring the large-scale complex networks [8, 9, 10, 11, 12, 13].

Ph.D. Candidate, Department of Modeling, Simulation and Visualiztion Engineering, Old Dominion University, 5115 Hampton Blvd., Norfolk, VA 23529. Email: xwang009@odu.edu.

Research Assistant Professor. Virginia Modeling, Simulation and Analysis Center, 1030 University Blvd., Suffolk, VA 23435. Email: ajcollin@odu.edu.

This dramatically increased visibility of SNA, however, is owed mainly to statistical physicists [14, 15, 16]. Instead of emphasizing agent heterogeneity, statistical physicists focus more on aggregate properties of large-scale networks, and highlight network's systematic regularities in spite of micro agent heterogeneity. Among many, Barabasi-Albert model (BA model) has attracted particular attention because of its novel perspective in revealing the mathematical properties of large-scale networks and its frequent appearance in a diverse range of network phenomena [17, 18].

Barabasi [17] argue that the vertex connectivities in large networks tend to follow a scale-free power-law distribution. The key underlying mechanism is that a vertex's probability to be connected is determined only by its relative position (i.e., connectiveness or "popularity") in the existing network. The formation of large networks is governed by this robust self-organizing mechanism that goes beyond the particulars of agents or individual systems. However, in many social networks considerable deviations from scale-free behaviors have been reported [19]. Numerous variants of BA model accordingly have been developed to reproduce the growth process of social networks, and most of them still share the very key instrumental assumption that a vertex's probability to be connected is determined primarily by its position (i.e., "popularity") in a given network.

In this study, we argue that another way to advance scholarly understanding of network formation is to "bring agent heterogeneity back in." As revealed in the SNA literature, particularly those of "homophily", people's decision to

establish social ties are not conditioned solely by instrumental calculations of the others' position in a network (e.g., "popularity"). If it is not more important, individuals are also motivated by intrinsic affection of joining the "like" (i.e., the preference to endogenous similarities). In other words, people are constantly weighting between popularity and proclivity in forming their social connections. The impact of this mixed preferential attachment, we argue, is particularly consequential on formation of social networks [20].

Although many empirical studies have confirmed the importance of agent heterogeneity at the vertex- and dyad- level [21], few studies, if any, have systematically explored its impacts on the aggregate mathematical properties of large-scale networks. In this study, we propose an integrative agent-based model (ABM) of heterogeneous attachment encompassing both instrumental calculation and intrinsic similarity. Particularly, we emphasize the ways in which agent-heterogeneity affects social network formation.

In three ways, this study contributes to current studies of network formation. First, by exploring the impacts agent heterogeneity, this studies highlights an important yet less examined mechanism in network formation, that is, intrinsic preferential attachment. This mechanism, we argue, becomes particularly important in the age of new media, in which individuals' capacity in homophilous sorting has been strongly boosted by ICTs [22, 23, 24]. Therefore, an investigation of the impacts of intrinsic preferential attachment can significantly enrich our understanding about large-scale social network. Second, by emphasizing both micro-mechanisms in governing dyad formation and macro mathematical properties of large-scale networks, we concur with [8, 413] that "the structure and the evolution of networks are inseparable." This study then provides an integrative perspective to understand the seemingly separate research enterprises of SNA, such as p^* models [25, 26] on the one hand and BA model on the other hand [17, 18, 27]. Third, joining many recent works [28, 29, 19], this study demonstrate that ABM, given its explicit emphasis on complexity and emergence, provides a promising perspective and a useful method to explore the dynamic evolution of large-scale social networks.

II. The Literature: Searching the Mechanisms of Network Dynamics

Social network is fundamentally a complex and emergent phenomenon, which raises great challenges in conceptualizing and theorizing network dynamics. Many scholars, recognizing the impacts of network structure, focus on how network position affects various socioeconomic outcomes. One of the most influential such works is Granovetter's "strength of weak ties" (SWT) theory [30, 31], in which weak, bridging ties are argued to be beneficial because of their potentials in introducing novel information. Burt [32] later refines the argument by differentiating between the benefits of bridging ties and the average strength of those ties. This in turn leads to Burt's conclusion that stronger ties can be beneficial than weak ties because they allow a greater flow of resources. More recently, Podolny [33] argues that network structure matters not only because it serves as "pipes" of resources, but also because it acts like "prisms," revealing important information about the inherent qualities of vertices (e.g., credibility).

A. *Understanding network dynamics*

Other scholars, rather than exploring the impacts of different network structures, focus on the network formation processes, ranging from the Erdos and Renyi's random graph model (ER model) to Watts and Strogatz's "small-world" model [34, 13]. However, as for large-scale complex networks, empirical results demonstrate that most of them are scale free, that is, their degree distribution follows a power law distribution (Redner 1998; Albert et al. 1999; Faloutsos 1999; Barabasi and Albert 1999; Broder et al. 2000; Newman 2001; Barabasi et al. 2001; Yook et al. 2001). BA model [17] then is introduced to describe this scale-free emergent mechanism. BA model

suggests that the growth of network size and preferential attachment are the necessary conditions for the emergence of scale-free networks (Albert and Barabasi 2002).

However, in many social networks, significant deviations from scale free behavior have been reported, and numerous complex network models have been proposed by updating the two key assumptions of BA model: dynamic growth and preferential attachment. It is believed that if the process that assembled the networks is captured accurately, it is possible to obtain the topology which is closer to real world networks. As for preferential attachment, Barabasi et al. (2001) and Newman (2001) estimate the functional form of preferential attachment via measuring the real world network data, including co-authorship network data, the scientific collaboration networks in physics and biology and the like. Krapivsky et al. (2000) propose a nonlinear preferential attachment network model and conclude that scale-free nature is not observed anymore when nonlinear preferential attachment mechanism is involved.

To accurately capture dynamic growth, Dorogovtsev et al. (2000) construct a attractiveness model that provides exact form of the distribution of incoming links of sites in the limit of large sizes of growing network. The degree is found to increase faster than the number of nodes in the system through some real large networks, such as WWW, against the linear growth assumed in BA model. Similarly, Dorogovtsev and Mendes (2000) examine the phenomenon of accelerated growth in the degree distribution. Klemm and Eguiluz (2002) propose a model that describes the dynamic growth process of the networks based on the finite memory of the nodes. .

B. Bring agent heterogeneity back in

However, Pujol et al. (2005) pointed out that the assumptions of these models usually lack sociology grounding. Wong et al. (2005) argued that many network models have not taken the advantages of sociological and psychological insights of how social networks may be formed. We also found it is problematic since it assumes all the nodes possess the same preference (instrumental preferential attachment) and overlooks the potential impacts of agent heterogeneity on network formation (intrinsic preferential attachment). When joining a real social network, people are not only driven by instrumental calculation of connecting with the popular, but also motivated by intrinsic affection of joining the like. In other words, people are constantly weighing between popularity and proclivity in forming their social connections. The impact of this mixed preferential attachment, we believe, is particularly consequential on such social networks as political communication. More importantly, we find the support to this assumption from the social theory: homophily.

Introduced by McPherson et al. (2001), homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people, and the similarity could be regarding to many types of personal characteristic positions, including gender, religion, social class, education and other intra-personal or behavioral characteristics. In fact, there are some models taking homophily into consideration, somehow not using the specific term but essentially the similar meaning. Robins et al. (2001) presented network models for social selection process. Although characteristic positions affecting the social relationship formation is concerned, it is broken between the local behavior and the global pattern. In other words, there is no analysis for the properties of large social networks. Newman and Girvan (2003) conducted a network model discussing the mechanism of assortative mixing, that is, the nodes with similar degree level like to link with each other. However, it actually is a special case of preferential attachment, albeit the similarity of nodes is concerned.

In this study, we propose an integrative model of preferential attachment encompassing both instrumental calculation and intrinsic similarity, which is a term transformed from homophily. Particularly, it emphasizes the ways in which agent-heterogeneity affects social network formation. Agent-based modeling is cho-

sen as the paradigm to conduct this study. This integrative approach, we believe can strongly advance our understanding about the formation of social networks.

C. The emergent agent-based modeling (ABM) approach

A review of ABM studies on networks reveals that there are two main directions. While many researchers interested in seeking the process of network formation, many others are working on exploring the diffusion under different networks, i.e., the transmission process in the context of varying network topologies. For example, Hamill and Gilbert (2009), in exploring network formation, argue that currently there is no such network models fit well with sociological observations of real social networks, and they provided an AB model based upon the social circle theory, which is different agents with unequal social reach, they create a wide variety of artificial social worlds labeled with properties of real world observed large scale networks. Similarly, Mitrovic and Tadic (Dynamics of bloggers' communities: bipartite networks from the empirical data and agent-based modeling 2012) conducted an analysis of the empirical data and ABM of the emotional behavior of users on the Web portals where the user interaction is mediated by posted comments. ABM here is used to simulate the dynamics and to capture the emergence of the emotional behaviors and communities. Gaston and Jardins (2005 agent-organized networks for dynamic team formation) provide the past findings of the structure of the artificial social network governing the agent interactions is strongly correlated with organizational performance. By the context of dynamic team formation, they proposed two strategies for agent-organized networks and evaluate their effectiveness for increasing organizational performance.

III. Model Hypothesis

In this study, we argue that another way to advance scholarly understanding of network formation is to bring agent heterogeneity back in. We landed our hypothesis on the grounding of existing sociological theory. Firstly, we consider the formation of complex social networks is driven by the human's intention. Consequently, mechanism of the network formation, we believe, should have its roots in the psychological construction of human being. Second, the complexity of network emergence is generated from the personal different choices according to the subtle insider of individuals. Hence, finding out the driven mechanism for making decisions of people is significantly important for clarifying the network formation simulation. We noticed that homophily, as a driven mechanism for social action, has been researched for a long time by some sociologists. In this section, we introduce the important driven mechanism theory in the first, and then list our entire hypothesis of our network model.

A. Heterogeneous attachment

Homophily, as a common observed phenomenon in the social world, has been learned and discussed for a long history. However, as a significantly important social factor, it is still not being taken into account when large-scale networks are simulated. Introduced by McPherson et al. (2001), homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people, and the similarity could be regarding to many types of personal characteristic positions, including gender, religion, social class, education and other intra-personal or behavioral characteristics. In fact, there are some models taking homophily into consideration, somehow not using the specific term but essentially the similar meaning. For instance, Robins et al. (2001) presented network models for social selection process. Although the fact that social relationship formation are affected by personal characteristics concerned, it is broken between the local behavior and the global pattern in terms of that the properties of large scale networks are not measured. In other words, there are no analyses for how does this important social rule affecting the behavior in the level of large scale social networks. Newman and Girvan (2003) conducted a network model dis-

cussing the mechanism of assortative mixing, which is, the nodes with similar degree level like to link with each other. However, it actually is a special case of preferential attachment, albeit the one among many similarities of agents is concerned.

In this study, we argue that homophily is another key driven mechanism, comparing to the mechanism of preferential attachment, for leading the social agents to make decisions for forming the different structures of the social networks. When joining a real social network, people are not only driven by instrumental calculation of connecting with the popular, but also motivated by intrinsic affection of joining the like. In other words, people are constantly weighting between popularity and proclivity in forming their social connections.

The impact of this mixed network formation is particularly consequential on such social networks as political communication. For instance, when people appear in a new community and start to build their network, the two endogenous driven mechanisms would lead people to build up their social networks. Under the extremely conditions, by following the preferential attachment, people only interested in linking with the popular people, so that the people have bigger potentials to expand their networks. By following the homophily, people would be more like to connect with the other people that have similar intrinsic properties with them, since they might be looking for a more comfortable social ambience or they even could gain more confidence from people owning the similar characteristic . Certainly, the latter is a human behavior factor, which is labeled as "intrinsic" intention to construct network in this study. We realize that in the real world, most people make decisions based on both mechanisms in different levels instead of in the extreme situations. In the following model design chapter, we will state the way of weighting for balancing these two mechanisms and the method of modeling homophily.

B. Hypothesis and heterogeneous model

The heterogeneous attachment model proposed in this study is rested on three key assumptions.

1) **Heterogeneity:** Vertices (i.e., agents) are intrinsically different from each other on certain aspects. All of the relevant characteristics of vertices are captured by a finite set of $c \geqslant 1$ types: $\{1, 2, ...c\}$. Based on this finite set of relevant characteristics, it is possible to construct C_i, representing the characteristic position of node i.
2) **Dynamic Growth:** The network continuously expands by the addition of new vertices. The network starts with a small number (n_0) of nodes, at each time step t, a new node with m edges that link the new node to m different nodes already present in the system.
3) **Heterogeneous Attachment:** The probability that two vertices are connected is jointly determined by the connectivity of the existing vertices and the intrinsic similarity between vertices. The joint probability that a new vertex at time step $t+1$ will be connected to vertex i depends on,

$$U = f(k_{it}, C_i, C_{t+1}) = \frac{\lambda k_{it}}{\sum_j k_{jt}} + (1-\lambda) \cdot g(C_i, C_{t+1}) \tag{1}$$

where λ is a weighted product of the instrumental preferential probability and intrinsic preferential probability. Connectivity of node i at time step t thus is k_{it}. The probability of a new node and a random existing node are connected for instrinsic purpose at time $t+1$ can be captured by $g(.)$, in which $g(C_i, C_{t+1})$ decreases as $C_{diff}(C_i, C_j)$ increases for $i \neq j$. C_{t+1} is the characteristic position of a new node entering the network at time step $t+1$, and C_i is that of node i already in the network, $i \in \{1, \ldots, t\}$.

In the original setting of the simulation, let $m = 1$, $n_0 = 2$ and $C_i \in \{\text{“Blue”}, \text{“Red”}\}$, and

$$g(C_{t+1}, C_i) = \begin{cases} \frac{1}{\mu N_d + N_s} & \text{for } C_{t+1} = C_i \\ \frac{\mu}{\mu N_d + N_s} & \text{for } C_{t+1} \neq C_i, \end{cases} \quad (2)$$

where N_s is the number of nodes $C_i = C_{t+1}$, N_d is the number of nodes $C_i \neq C_{t+1}$.

IV. Walking rules of agents in NetLogo

In order to illustrate and model heterogeneous attachment, this study uses different colors to denote the different attributes of agents as stated in Hypothesis 1. Specifically, for the purpose of simplicity there are two types of agents in the system: blue agents and red agents. We model Hypothesis 2 of by allowing the size of agents to increase at each tick, a default time unit in NetLogo. It should be noted that this study focus primarily on the topology of social networks obtained on the final stage. Therefore, we assume simple dynamic growth of agents in this model. The network will stop to grow when there are 10000 agents in the network. As implied in the formula (1) of Hypothesis 3, when $\lambda = 1$ a purely rationality-driven social network (i.e., a classical scale free network) is expected to emerge, and when $\lambda = 0$, a value-driven social network is expected to be generated. Agents in the model would take the same color agents into their homophily consideration. The same process will repeat for 30 times for each λ, $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$.

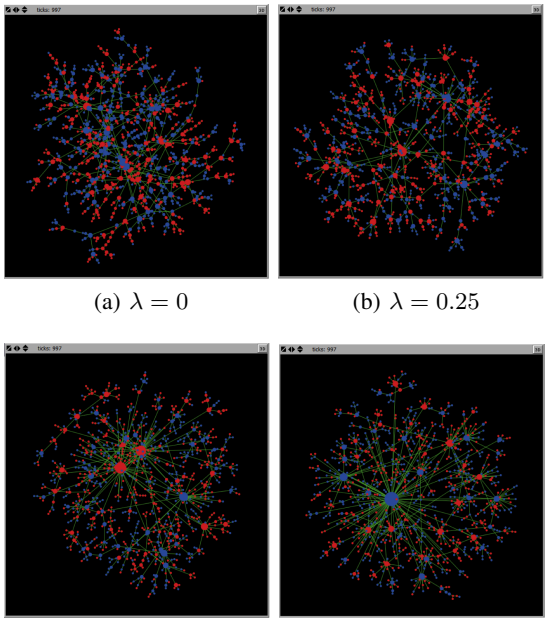
More specifically, the simulation process can be described as the following steps:

1) Start with two connected nodes with random values (red/blue color).
2) A new node with a random values (red/blue color) is starting to consider about joining in the networks.
3) The nodes existent in network (for first run, the original two nodes) are in the choosing queue.
4) The incoming node is weighing between popularity and proclivity. It choose one node from the queue to connect by calculating the probability given in Formula (1). The node with higher probability in the existing network is more likely to be connected.
5) The fourth node is entering the network, making three existing nodes in the waiting queue. The probabilities are calculated for each existent node, which in turn are used to decide which node is to be connected.
6) Loop from step 3.
7) Each run of simulation stops when there are 10000 nodes in the network.
8) Simulation stops when the model runs 30 times for each λ, $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$.

A real world instance can help illustrate the model procedure discussed above. A person newly moved into a community has to build up relationships with others. At the point of considering who is a suitable “friend,” he is guided by two heuristics: “Is he or she popular?” and “Is he or she a person more like me?” After weighing these two considerations, the new comer makes a choice and chooses one to link. With the expansion of this community, this new comer becomes an existent member and thus an candidate to be evaluated by more new comers.

We are interested in exploring the ways in which the inherent characteristics of individuals, that is, different predispositions of heterogeneous people, affect the structure of scale free networks. Particularly, this study intends to examine if there is any turning point or linking point for generating network topologies differently. Theoretically, this study help reveal how network formation process affected by including the assumption of heterogeneity for autonomous agents. To do so, we simulate the formation of networks model with different values of “λ”, where “λ” indicates to what extent people emphasize “more like me” in choosing their friends. Specifically, our experiment includes five different values of λ variable:

$\lambda = 0$	People only concern about “Are you more like me?” or “Are we in the same party?”
$\lambda = 0.25$	People concern more about “Are



(a) $\lambda = 0$ (b) $\lambda = 0.25$

(c) $\lambda = 0.75$ (d) $\lambda = 1$

Figure 1: Visualization of simulated networks

you in my side", and also care about "Do you have more links" a little bit.

$\lambda = 0.5$ People concern these two parameters in the same level.

$\lambda = 0.75$ People concern more about "Do you have more links", and also care about "Are you in my side" a little bit.

$\lambda = 1$ People only care about "How many links do you have?"

After simulating our network model in NetLogo, we obtain networks with different values of λ. The visualization of the simulated networks can be generated as Figure 1.

Visually and roughly, we can observe a pattern across the five graphs as the value of λ increases. When the value of λ increases, there are more super nodes (nodes with a large number of links) in the system. When λ value decreases, nodes are connected more evenly and there are almost no observable super nodes. In other words, we could explain this observation as that when people make decisions according to the preferential attachment, there would be many more monopolies who own lots social resources in the emergent social networks. When people make decisions upon the heterogeneous attachment, the social resources may evenly distributed. However, this is only an observation and a rough inference based on the model visualization we have so far. Introduced by Barabasi (1999), degree distribution of preferential attachment follows a power law distribution. Degree distribution refers to the number of linker of each node. Instinctively, degree distribution of heterogeneous attachment should follow exponential distribution. In next section, we conducted the statistical analysis on the degree distribution outputted from our agent-based model.

V. Methodology

Following statistical analysis techniques on the power law distribution provided by [18], we conducted an analysis on the degree distribution of the data generated from the simulation. The main goal of this statistical analysis is to learn how the degree distribution changes according to the different λ values, furthermore, to test our hypotheses that when $\lambda = 1$, we are reasonable to conclude that the degree distribution follows a power law distribution and when $\lambda = 0$, the exponential distribution is the good one to describe the degree distribution.

A. Definitions relating to power law distribution

Mathematically, a quantity x obeys a power law if it is drawn from a probability distribution:

$$p(x) \propto x^{-\alpha} \tag{3}$$

Introduced by Watts (2004), the probability of a randomly chosen node having degree x decays like a power of x, where the exponent α, typically measure in the range of $2 < \alpha < 3$, determines the rate of decay (smaller α implies slower decay, hence a more skewed distribution). A distinguishing feature of power-law distributions is that when plotted on a double

logarithmic scale, a power law appears as a straight line with negative slope α. Argued by Clauset et. al (2009), few empirical phenomena obey power laws for all values of x. Usually, the power law applies only for the values greater than some minimum $x_{min.}$ Basiclly, power law distribution have two different settings: continuous distributions with the continuous real number and discrete distributions with discrete set of positive integers. Since the data of our model are positive integers, the probability distribution should follow the form of:

$$p(x) = \Pr(X = x) = Cx^{-\alpha}. \tag{4}$$

This density function diverges at $x = 0$, implying a lower bound $x_{min} > 0$ to the power law behavior. The normalizing constant can be calculated as follow equations:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} \tag{5}$$

where

$$\zeta(\alpha, x_{min}) = \sum_{n=0}^{\infty} (n + x_{min})^{-\alpha}. \tag{6}$$

B. Analyzing power law distributed data

Based on th power law distribution, provided by Clauset et. al (2009), the approach is used for analyzing our data generated from the simulation to test how the λ value affects the degree distribution, so as well as the system behavior of the network formation process. We follow the below steps:

1. Estimate the parameters of power law distribution: $x_{min}$and scaling parameter λ.

2. Calculate the goodness-of-fit between our data and the power law. If the p-value is greater than 0.1, the power law is a plausible hypothesis for the data, otherwise it is rejected.

3. Compare the power law with alternative hypothesis. Here, a likelihood ratio test method is approached. For each alternative, if the calculated likelihood ratio is significantly different from zero, then its sign indicates whether or not the alternative is favored over the power law distribution.

1) Estimating parameters: There are two parameters need to be estimated in power law model: $x_{min}$and scaling parameter α. The method of estimating $x_{min}$is introduced by Clauset, Young and Gleditsch (2007) and it suits to both discrete and continuous data. The fundamental idea is: the lower bound $\hat{x}_{min}$should make the probability distributions of the measured data which are above this lower bound and the best-fit power law model as similar as possible. They choose Kolmogorov-Smirnov or KS statistic as a measurement of quantifying the distance between the two probability distributions, which indicates the maximum distance between the cumulative distribution functions (CDFs) of the measured data and the fitted model:

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \tag{7}$$

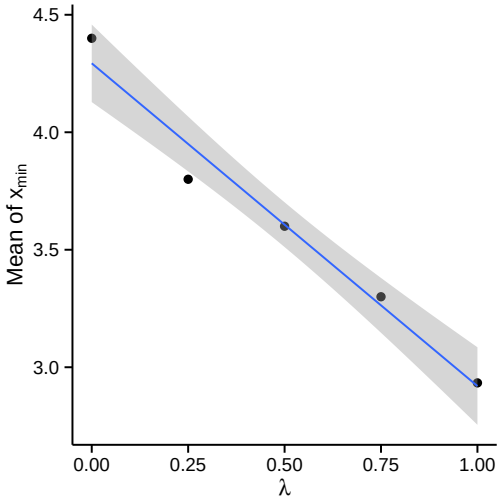
where $S(x)$is the CDF of the simulated data for the observations with the value greater than $\hat{x}_{min}$, and the $P(x)$is the CDF for the power law model that best fits the data in the region $x \geq \hat{x}_{min}$. Hence the $\hat{x}_{min}$is selected to be the x_{min}that minimizes the value of D.

Referring to the α estimating, method of maximum likelihood provably gives the accurate parameters estimates with the limit of large sample size. Assuming the data are drawn from a distribution following a power law for $x \geq x_{min}$, the scaling parameter estimation under discrete case can be derived through maximum likelihood estimators (MLEs) as:

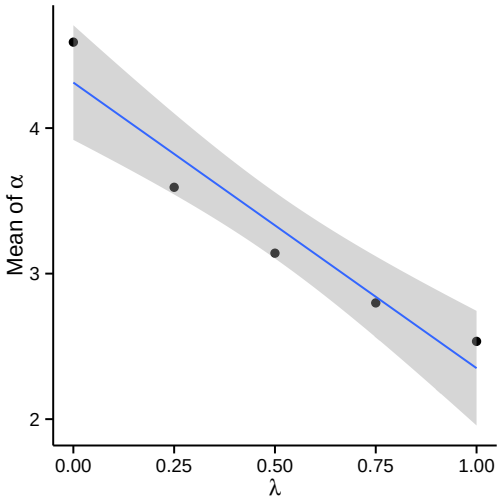
$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]^{-1} \tag{8}$$

where x_i, $i = 1, ..., n$, are the observed values of x such that $x_i \geq x_{min}$.

Following this procedure, we estimated the parameters for the data from the simulation. According to the experiments of our simulation, with 5 different λvalues, we estimated 30 pairs of parameters based on 10000 data points under each λvalues. We did a short summary based on the estimates we got so far and accordingly provide the pre-conclusions based on the data of estimates:



(a) x_{min}



(b) α

Figure 2: Relationship between λ and the mean of x_{min} and α

Observed from the graphs above, we can conclude that the means of $\alpha$and x_{min} are significantly changed according to the different λvalues. When λvalue close and equal to 1, the mean of αvalue falls in the range of $[2, 3]$, which is the similar to the αvalue of real world power law distributed data. As well as the change of x_{min}value, it decreases as the value of λ increase. We may conclude here that

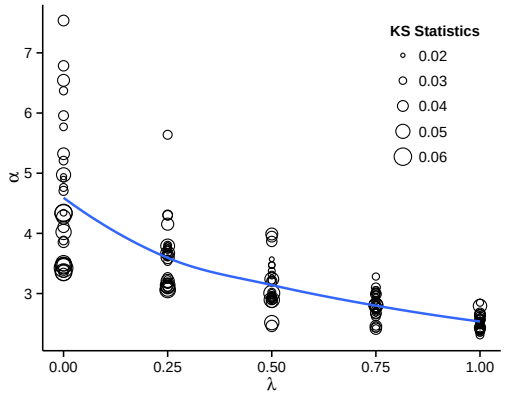


Figure 3: Relationship between λ and KS statistics

as λvalue increase, indicated by the decreased lower bound of x_{min}, the more data points are included and used in testing the power law distribution for next step. The second graph illustrates the 30 KS statistics under each λvalue. The area of the circle indicates the value of KS statistics. We see that the KS statistics are significantly affected by the change of λvalues as well. When λcloses to and equals to 1, the KS statistics values turn to consistent and convergent. Conversely, when λcloses to and equals to 0, the KS statistics values turn to uneven and sparse. To test our hypothesis, we still need more mathematical evidence.

2) Testing the power law hypothesis : In last section, the parameters of the power law are estimated based on our simulated data, this step provides the detail of telling if the fit is a good match to the data. The techniqual details provided by Clauset et. al are as following:

1.We fit our simulated data to the power law model as described in last section and KS statistics for the fit are calculated.

2.We generate a large number of power law distributed synthetic data sets with scaling parameter $\alpha$and lower bound x_{min}equal to those of the distribution that best fits the observed data. Here, each synthetic data set individually be fitted to its own power law model and accordingly the KS statistic is counted for

each one relative to its own model.

3. Count the fraction of the time that the resulting statistic is larger than the value for the simulated data. This fraction is our p-value. Since we have 30 runs under each λ value, we generated 30 p-value under each λ value.

3) Comparing with the alternative distributions: The step in last section provide a test if our simulated data is possibly drawn from a power law distribution. However, is it still possible that another distribution, such as an exponential or a log-normal, might give a fit as good or better? In this section, we need to tell if power law distribution is the one for our simulated data, other than the other alternative distributions. Given by the method mentioned in the last two sections, we only need to run through the whole process again for different distribution candidates. Relying on the p-values, make the judgment of accepting or rejecting our hypothesis.

VI. Conclusions

We build up our model, run the simulation and conduct the data analysis through the whole process above. We find out the results are consistent with our hypothesis:

1. How agents balance between intrinsic goods and instrumental goods strongly structures the formation of social network.

2. Mathematically, when $\lambda = 1$, the degree distribution can be explained using power law distribution. When $\lambda = 0$, the degree distribution can be described in exponential distribution.

References

[1] Peter J. Carrington, John Scott, and Stanley Wasserman, Eds., *Models and Methods in Social Network Analysis*, Cambridge University Press, Cambridge and New York, 2005.

[2] David Knoke, *Political Networks: The Structural Perspective*, Cambridge University Press, Cambridge and New York, 1990.

[3] John Scott, *Social Network Analysis: A Handbook*, Sage, London, 2nd edition, 2000.

[4] Stanley Wasserman and Katherine Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge and New York, 1994.

[5] Paul F. Lazarsfeld and Robert K. Merton, "Friendship as a social process: A substantive and methodological analysis", in *Freedom and Control in Modern Society*, Morroe Berger and Theodore Abel, Eds., pp. 18–66. Van Nostrand, New York, 1954.

[6] Linton C. Freeman, "Some antecedents of social network analysis", *Connections*, vol. 19, no. 1, pp. 39–42, 1996.

[7] Miller McPherson, Lynn Smith-Lovin, and James M. Cook, "Birds of a feather: Homophily in social networks", *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[8] Albert-László Barabási, "Scale-free networks: A decade and beyond", *Science*, vol. 325, no. 5939, pp. 412–413, 2009.

[9] Guido Caldarelli, *Scale-Free Networks: Complex Webs in Nature and Technology*, Oxford University Press, Oxford and New York, 2007.

[10] Karen Heyman, "Making connections", *Science*, vol. 313, no. 5787, pp. 604–606, 2006.

[11] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne, "Computational social science", *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[12] Mark Newman, Albert-László Barabási, and Duncan J. Watts, *The Structure and Dynamics of Networks*, Princeton University Press, Princeton, NJ, 2006.

[13] Duncan J. Watts, "The 'new' sicence of networks", *Annual Review of Sociology*, vol. 30, pp. 243–270, 2004.

[14] Ajith Abraham, Aboul-Ella Hassanien, and Vaclav Snasel, *Computational Social Network Analysis: Trends, Tools and Research Advances*, Springer, London and New York, 2010.

[15] Ulrik Brandes and Thomas Erlebach, *Network Analysis: Methodological Foundations*, Springer, New York, 2005.

[16] Réka Albert and Albert-László Barabási, "Statistical mechanics of complex networks", *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[17] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks", *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[18] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman, "Power-law distributions in empirical data", *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[19] Amir Hossein Shirazi, Ali Namaki, Amir Ahmad Roohi, and Gholam Reza Jafari, "Transparency effect in the emergence of monopolies in social networks", *Journal of Artificial Societies and Social Simulation*, vol. 16, no. 1, pp. 1–9, 2013.

[20] Sinan Aral, Lev Muchnik, and Arun Sundararajan, "Engineering social contagions: Optimal network seeding in the presence of homophily", *Network Science*, vol. 1, no. 2, pp. 125–153, 2013.

[21] Tom A. B. Snijders, Gerhard G. van de Bunt, and Christian E. G. Steglich, "Introduction to stochastic actor-based models for network dynamics", *Social Networks*, vol. 32, no. 1, pp. 44–60, 2010.

[22] Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, Yale University Press, New Haven, CT, 2006.

[23] Paul DiMaggio, Eszter Hargittai, W. Russell Neuman, and John P. Robinson, "Social implications of the internet", *Annual Review of Sociology*, vol. 27, pp. 307–336, 2001.

[24] Kevin Lewis, Marco Gonzalez, and Jason Kaufman, "Social selection and peer influence in an online social network", *Proceedings of the National Academy of Sciences*, vol. 109, no. 1, pp. 68–72, 2012.

[25] Steven M. Goodreau, "Advances in exponential random graph (p*) models applied to a large social network", *Social Networks*, vol. 29, no. 2, pp. 231–248, 2007.

[26] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher, "An introduction to exponential random graph (p*) models for social networks", *Social Networks*, vol. 29, no. 2, pp. 173–191, 2007.

[27] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen, "Problems with fitting to the power-law distribution", *The European Physical Journal B*, vol. 41, no. 2, pp. 255–258, 2004.

[28] Lynne Hamill and Nigel Gilbert, "Social circles: A simple structure for agent-based social network models", *Journal of Artificial Societies and Social Simulation*, vol. 12, no. 2, 2009.

[29] John H. Miller and Scott E. Page, *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*, Princeton University Press, Princeton and Oxford, 2007.

[30] Mark S. Granovetter, "The strength of weak ties", *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[31] Mark S. Granovetter, "The strength of weak ties: A network theory revisited", *Sociological Theory*, vol. 1, pp. 201, 1983.

[32] Ronald S. Burt, "The contingent value of social capital", *Administrative Science Quarterly*, vol. 42, no. 2, pp. 339–365, 1997.

[33] Joel M. Podolny, "Networks as the pipes and prisms of the market", *American Journal of Sociology*, vol. 107, no. 1, pp. 33–60, 2001.

[34] Duncan J. Watts and Steven H. Strogatz, "Collective dynamics of 'small-world' networks", *nature*, vol. 393, no. 6684, pp. 440–442, 1998.