OLD DOMINION UNIVERSITY
IDEA FUSION

Virginia Modeling, Analysis and Simulation Center

Modeling, Simulation & Visualization
2013 Student Capstone Conference

Final Proceedings

April 11, 2013
Old Dominion University

2013 Student Capstone Conference Final Proceedings Introduction

2013 marks the seventh 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 44 research works that were presented on April 11, 2013 at the conference.

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

- Science & Engineering
- Homeland Security/Military M&S
- Medical M&S
- Gaming & VR
- Training & Education
- Transportation
- Business & Industry

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

Those recipients are listed below by track.

Science & Engineering:

Best Paper: Jan Nalaskowski - International Studies -ODU
Best Presentation: Sirisha Mushti - Department of Mathematics and Statistics -ODU

Homeland Security & Military:

Best Paper: Nicholas Wright - United States Military Academy
Best Presentation: Christopher Lynch - VMASC/MSVE - ODU

Medical & Health Care Simulation:

Best Paper: Andrew McKnight - Computer Science Department -ODU
Best Presentation: Erik Prytz - Department of Psychology - ODU

Gaming & Virtual Reality:

Best Paper: Gary Lawson MSVE - ODU
Best Presentation: Carmelo Padrino-Barrios RDH-ODU

Training & Education:

Best Paper: Umama Ahmed MSVE - ODU
Best Presentation: Rebecca Law - International Studies - ODU

Transportation Track:

Best Paper: Khairul Anuar -Civil Engineering Department - ODU
Best Presentation: Terra Elzie VMASC/MSVE - ODU

Business & Industry:

Best Paper: Daniele Vernon-Bido VMASC/MSVE - ODU
Best Presentation: Daniele Vernon-Bido VMASC/MSVE - ODU

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.

For the 2013 Student Capstone Conference, The Gene Newman Award went to: ***Andrew McKnight***, from the Computer Science Department for his paper entitled '*Estimating Lower Bounds on the Length of Protein Polymer Chain Segments using Robot Motion Planning*', this work was co-authored by Jing He, Nikos Chrisochoides and Andrey Chernikov.

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.

Science and Engineering

VMASC Track Chair: Dr. Bridget Giles
MSVE Track Chair: Dr. Masha Sosonkina

Modeling Two Dimensional Molecular Couette Flows in a Nano-Scale Channel
Author(s): Wei Li, Zhaoli Guo, and Li-Shi Luo

Experience of Using AHP for ASV Design Improvement
Author(s): Bradley Lesher, Christopher Johnson, Timothy Hahn, and Michael LaPuma

Simulation of Dynamic Structures of Active Nematic Nano Particle
Author(s): Panon Phuworawong, and Ruhai Zhou

Measuring Success of Separatists' Demands: Development of the Tool
Author(s): Jan Nalaskowski

Modeling and Analysis of Continuous Longitudinal Data
Author(s): Sirisha Mushti, and N. Rao Chaganty

Describing the Role of Calibration in System Dynamics Models: Some Issues and Answers
Author(s): Ange-Lionel Toba

An Analysis of Various GPU Implementations of Saint-Venant Shallow Water Equations
Author(s): Joseph C. Miller III

Simulation on the Clouds
Author(s): Hamdi Kavak

Exploring the M&S Body of Knowledge
Author(s): Olcay Sahin

Autocorrelation Functions Applied to the Benthic Community of the Chesapeake Bay Suggest a Lack of Seasonality
Author(s): Kevin Byron, and Daniel Dauer

Modeling two dimensional molecular Couette flows in a nano-scale channel*

Wei Li
Department of Mathematics & Statistics and
Center for Computational Sciences,
Old Dominion University, Norfolk, Virginia 23539, USA

Zhaoli Guo
State Key Laboratory of Coal Combustion,
Huazhong University of Science & Technolgy, Wuhan 430074, China

Li-Shi Luo
Department of Mathematics & Statistics and
Center for Computational Sciences,
Old Dominion University, Norfolk, Virginia 23539, USA

Abstract

The burgeoning development of nano technology has intrigued numerous researches on nano scale flows. The study of molecular gaseous flows in macro or nano scale channels have been widely applied in realms like micro-reactors [1], nanopumps [2], and other lab on a chip micro-fluids [3]. In molecular gaseous flows in microfluidics, detailed molecule-surface interaction that cannot be considered in continuum theory are usually neglected in kinetic theory. However the forces are dominant effect to produce near surface physical phenomenon. Consequently, molecular dynamics (MD) becomes an indispensable computational tool to study molecular gaseous flows in micro or nano scales. There have been studies using MD to investigate

*The paper based on this research has been submitted to Physical Review Letters and is under review.

various gaseous flows, such as Couette flow and Poiseuille flow [4] [6]. The MD simulations reveal microscopic nature of molecular flows in microfluidics which are hard to obtain by other means. However, because of its demanding requirement on computational resources, MD is not yet a viable tool for realist problems of engineering application. There is a pressing need to develop effective and efficient computational models to simulate molecular gaseous flows in microfluidics.

The present study is a trial to expedite the simulation of molecular flows. We propose an approach based on kinetic and continuum theory to model the molecular flows in micro or nano-scale channels, in which the wall-molecule interaction is the dominant effect. This approach is used to reproduce the density and velocity of the molecular Couette flow in two dimensions with the modified Knudsen number $k = 10.0$ and $k = 1.0$ obtained by MD simualtions [5]. Specifically, we propose the concept of effective radial distribution functions according to kinetic theory. By observation, on one hand the averaged density profile from MD simulation can be approximated by a certain type effective radial distribution function; on the other hand, the averaged velocity profile can be approximated by the product of another type of effective radial distribution function and the velocity from the linearized Boltzmann equation. Via determining the parameters in our density and velocity model, we procure smooth curves of the modeled density and velocity profile of the underlying flow. This smooth profiles enable us to evaluate the corresponding effective viscosity profile of the flow by using continuum theory, which completes the modeling. To validate our model, We implement lattice Boltzmann equation (LBE) simulation with the modeled effective viscosity profile and the same wall-molecule force field used in MD simulation. The LBE simulation demonstrates the proposed approach can accurately reproduce the phenomena due to the molecule-wall interaction dominated in the MD simulation. Moveover, our model accelerates the simulation by at least two orders of magnitude in term of the computational time.

References

[1] J. Kobayashi, Y. Mori, K. Okamoto, R. Akiyama, M. M. Ueno, T. Kitamori, and S. Kobayashi, Science **304**, 1305 (2004)

[2] C. Liu and Z. Li, Phys. Rev. Lett. **105**, 174501 (2010)

[3] A. Gunther, S. A. Klan, M. Thalmann, F. Trachsel, and K. F. Jensen, Lab Chip **4**, 278 (2004)

[4] D. K. Bhattacharya and G. C. Lie, Phys. Rev. Lett. **62**, 897 (1989)

[5] M. Barisik, B. Kim, and A. Beskok, Commun. Comput. Phys. **7**, 977 (2010)

[6] M. Barisik and A. Beskok, Microfluid. Nanofluid. **11**, 611 (2011)

Simulation of Dynamic Structures of Active Nematic Nano Particle

Panon Phuworawong, Ruhai Zhou
Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529

I. Introduction

We consider an ensemble of rod-shaped, rigid, active nematic nano particles (or active liquid crystal polymers). These particles are driven in the suspension either by their own biological/chemical forces or external electric/magnetic fields. Hydrodynamic interactions introduce complex spatio-temperal structures with strong fluctuations and long range correlations. There has been much research effort in the past decade to understand the dynamics, interactions and structures of active nematic suspensions which is a key to produce high performance active materials with desired properties. The goal in this research is to numerically study the behavior of local particle concentration, polarity and nematic director, and the velocity field, as well as their correlations.

II. Problem

Recently a kinetic theory is proposed for dilute and semidilute active liquid crystal polymer solutions where the concentration is low [1]. Each particle is described by its center-of-mass position $\mathbf{x}$ and axis of symmetry $\mathbf{m}$. The force generated by the active rods is assumed to be along the direction of $\mathbf{m}$. Let $f(\mathbf{x},\mathbf{m},t)$ be the active nematic particle distribution density. The dimensionless kinetic model consists of the Smoluchowski equation for f and the Navier-Stokes equation:

$$\begin{array}{l}\frac{\partial f}{\partial t}+\nabla\cdot\left(\left(\mathbf{v}+U_0\left(\bar{\alpha}\mathbf{m}+\sqrt{1-\bar{\alpha}^2}\mathbf{m}^{\perp}\right)\right)f\right)= \\ \quad \nabla\cdot D_s^*\left(\nabla f+f\nabla U\right)+\frac{1}{De}\mathcal{R}\cdot\left(\mathcal{R}f+f\mathcal{R}U\right) \\ \quad -\mathcal{R}\cdot\left(\mathbf{m}\times\dot{\mathbf{m}}\,f\right), \\ \frac{d\mathbf{v}}{dt}=\nabla\cdot\left(-p\mathbf{I}+\tau+\tau_a\right)-\langle\nabla\mu\rangle\,, \qquad \nabla\cdot\mathbf{v}=0\end{array} \tag{1}$$

where $\tau_a = G\zeta_a(\mathbf{M}-\frac{\phi\mathbf{I}}{3})$ is the active stress generated by the active force, $\mathbf{M}$ is the second moment of f, τ is other extra stress, $\mathcal{R}$ is rotational gradient operator, $\dot{\mathbf{m}}$ is Jeffery orbit, and potential

$$U = N_1\langle 1\rangle-\gamma\langle\mathbf{m}\rangle\cdot\mathbf{m}-\frac{3N}{2}\langle\mathbf{mm}\rangle:\mathbf{mm}$$

where N_1,γ,N are strength of space in homogeneity, polar, and nematic interactions respectively. We refer the reader to [1] for other symbols in the equation.

In this research, the model is considered in two dimensional physical space where the orientation can be written as $\mathbf{m} = (\cos\varphi,\sin\varphi)^{\mathrm{T}}$.

III. Methods

We approximate the active nematic particle distribution density by truncated Fourier series

$$f(\mathbf{x},\mathbf{m},t)\approx\sum_{k=0}^{K}\left(a_k(\mathbf{x},t)\sin(k\varphi)+b_k(\mathbf{x},t)\cos(k\varphi)\right) \tag{2}$$

This research impose $K=20$ which correspond to a system of 41 nonlinear PDEs in physical space $\mathbf{x}$ and time t that are strongly related to the hydrodynamic variables and flow equations. Then, we solve the system numerically. The discretization for the spatial variables follows a standard finite difference method. For the time integration of Smoluchowski equation, we use the linearly semi-implicit scheme

$$\begin{array}{l}\left(\mathbf{I}-D_s^*\Delta t\nabla^2-\frac{\Delta t}{De}\mathcal{R}\cdot\mathcal{R}\right)f^{n+1}= \\ f^n-\Delta t\nabla\cdot\left(\left(\mathbf{v}^n+U_0\left(\bar{\alpha}\mathbf{m}+\sqrt{1-\bar{\alpha}^2}\mathbf{m}^{\perp}\right)\right)f^n\right) \\ +D_s^*\Delta t\nabla\cdot\left(f^n\nabla U^n\right)+\frac{\Delta t}{De}\mathcal{R}\cdot\left(f^n\mathcal{R}U^n\right) \\ -\mathcal{R}\cdot\left(\mathbf{m}\times\dot{\mathbf{m}}f^n\right)\end{array} \tag{3}$$

The Navier-Stoke equation is solved by a projection method with staggered grid [2], [3], [4].

IV. Results

The simulations are conducted on square and rectangle domains by varying nematic strength N and active parameter ζ_a. In the case of pusher, $\zeta_a < 0$, we observe various interesting spatio-temperal structures.

For $N=1,\zeta_a=-4$, the simulation is performed in a unit square domain. Figure 1(a)(b)(c) show some snapshots of concentration, velocity, polarity and nematic director fields at one time. We observe strong fluctuations of the local concentration (indicated by the background color in Fig. 1a). There are four quadrants in the domain with two circles and two hyperbolic points. At the circles, the concentration takes the minimum value and the nematic orientational order (Fig. 1c) is weak, while at the hyperbolic points, the concentration takes the maximum value and the nematic orientational order is strong. Yet the flow field (Fig. 1b) is weak at both circles and hyperbolic points, it is at its peak value in the middle of the transition from one circle to another. Figure 1d shows the correlation between particle polarity and fluid velocity fields along the simulation time. Most of the time in the period, the polarity director positively correlates with the flow (indicated by the value 1 of the cosine angle between these two fields). But reversal occurs when the flow is weak and the polar

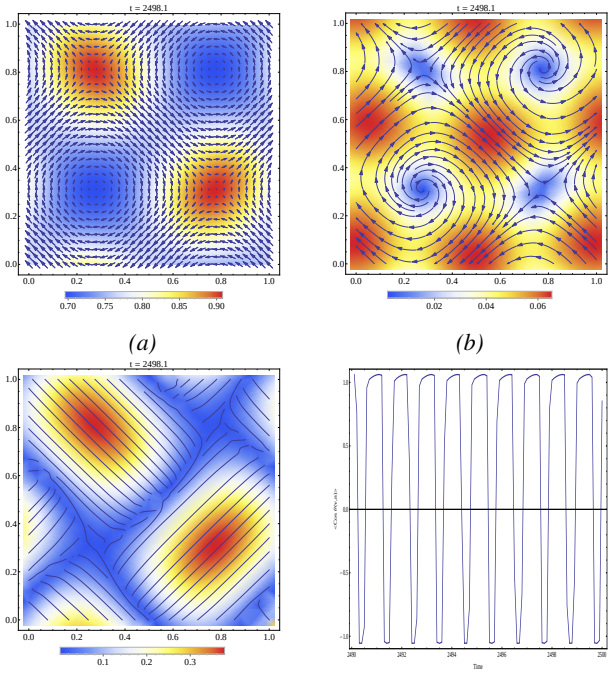


Fig. 1: The simulation on a 1×1 *square domain with* $N = 1$*, and* $\zeta_a = -4$*. (a) Velocity field superimposed to the density plot of concentration. (b) Polarity director. (c) Nematic orientation. (d) The domain averaged cosine angle between the particle polarity director and the fluid velocity.*

director goes against the flow direction (indicated by the value -1 of the cosine angle).

Another simulation is performed in a rectangular region for $N = 0.5$ and $\zeta_a = -2.5$. Strongly oscillatory spatio-temporal structure is also observed. Figure 2 shows the snapshots of the flow field (superimposed to the local concentration) and the polarity director field. There are two shear layers (roughly the top half and the bottom half) with opposite flow and polarity directions. At the concentration valley, the flow field is strong, while at the concentration peak, the velocity is almost quiescent. A movie of this structure shows rapid flow reversal when these two layers change direction.

V. Conclusion

This research conducts 2D numerical simulations of active nematic polymers using the kinetic model [1]. Local concentration, velocity field, polar direction, and nematic orientational order are explored in the dilute regime where the stable passive isotopic phase is driven out of the equilibrium. Due to the active stress, rapid concentration fluctuation is observed in the physical domain for some concentration and active parameters. Polarity and the orientational directors are closely correlated with the flow field in the period of structure oscillations. These findings are in agreement with reports in [5] when a linear model is used.

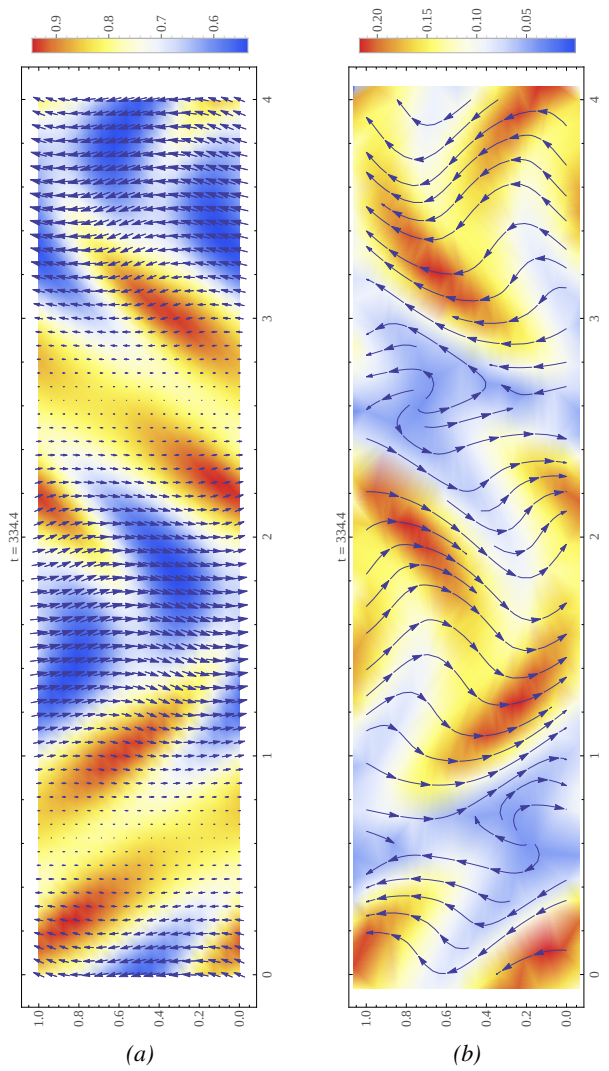


Fig. 2: The polarization direction on a 4×1 *rectangle domain with* $N = 0.5$*, and* $\zeta_a = -2.5$*. (a) Velocity field superimposed to the density plot of the concentration. (b) Polarity director.*

References

[1] M.G. Forest, Q. Wang, R. Zhou, *Kinetic Theory and Simulations of Active Nematic Polymers*, Soft Matter (accepted 2013)

[2] A. J. Chorin, *Numerical solution of the Navier-Stokes equations*, Math. Comp., 22 (1968), 745762.

[3] C. Hirt, B. Nichols, N. Romero, *A Numerical Solution Algorithm for Transient Fluid Flows*, Technical Report (1975)

[4] C. M. Tome, S. McKee, *GENSMAC: A Computational marker and cell method for free surface flows in general domains*, J. Comput. Phys.,110 (1994), 171-186

[5] D. Saintillan and M. J. Shelly, *Instabilities, Pattern Formation and Mixing in Active Suspensions*, Phys. Fluids 20, 123304 (2008)

This work is supported by NSF grant DMS-0908409 and Modeling & Simulation Scholarships from Old Dominion University.

Measuring Success of Separatists' Demands: Development of the Tool

Jan Nalaskowski

***Abstract*—The paper examines question of independence and separatist demands formulated by regional entities. It introduces the Monte Carlo model and simulation in order to check under what conditions a mother-state is more prone to accept these demands. The system is described with the real world data, serving as a rationale for random variables generators. Using Selectorate Theory as a behavior mechanism, the model provides a nucleus which can be further developed as an analytical and visual tool.**

***Index Terms*—Authoritarianism and democracy, Selectorate Theory, separatist and independence movements, public and private goods provision.**

I. Introduction

CONCEPTUALLY, this model aims to catch and operationalize an important, real world phenomenon. Independence and separatist movements are both significant and difficult to assess. Due to their nature, social sciences usually employ descriptive and quantitative analysis of these processes, but resulting conclusions are often ambiguous. Some point out domestic explanations, other prefer to acknowledge global, international tendencies.

The Selectorate Theory offers a parsimonious, quantitative analysis of internal policy-making and foreign outcomes. The logic of this model is to introduce this theory and project it to the sphere of independence and separatist movements.

Despite the fact that explanations are by definition internal, the model points out the future need to implement international factors. Therefore, the main goal of this model is to introduce a parsimonious, quantitative framework in inherently descriptive sphere of independence and separatist tendencies. It is to become a nucleus for further improvements. Data gathering, case studies, game theoretic component and implementation of international factors will make this model the useful tool for scholars and policy-makers.

First, the simuland is described. Then, the Selectorate Theory is introduced. Both conceptual and operational models are thoroughly explained. The results of Monte Carlo simulation are discussed and model's implications assessed. Verification and validation proceedings are sketched, revealing certain limitations and the need of further development.

Manuscript submitted on March 21, 2013.

Jan Nalaskowski is with Graduate Program in International Studies, Old Dominion University, Norfolk, VA 23529 USA (e-mail: jnalasko@odu.edu).

II. Simuland

Many nation states consist of regions which reveal certain ambitions for greater autonomy. This model acknowledges these tendencies without paying attention to their motives. The ambition itself is assumed to be present and it is mother-state's reaction towards it that is important. Ultimately, a mother-state decides whether to grant independence to its region or not. It is this particular moment that model tries to catch. Further implications in the form of civil wars or international intervention, as for example in cases of Abkhazia, South Ossetia or even Kosovo, are beyond the scope of this model. Therefore the question is – will mother-state itself grant independence to a region or not. The Selectorate Theory is further explained, but the main idea is that every state's political decision derives from its leader's willingness to remain in office. This analytical parsimony makes it possible to explain state's attitude towards independence ambitions of a region from internal level.

The model consists of many random generators which were created basing on the real-world data. Their purpose is to provide simulation results to empower Selectorate Theory with synthetic, yet vital variables. The model is innovative but its main role is to execute multiple operations, saving time and energy of researchers. It should be coupled with human intelligence and problem solving abilities to become the usable tool.

In short, simuland here is the real-world phenomenon of ambition for independence, revealed by many regions being parts of certain mother-states, which then express their answers to these requests. Some answers are positive, like the cases of Greenland or Faroe Islands, but most are negative, for example Catalonia or Tibet. All these regions have certain characteristics which are included into this model. Population variable is a fundament on which modeling, simulation and visualization is based. Assessment of country's wealth is proposed as a vital factor for Selectorate Theory's explanation. Its link with the outcomes has to be further developed but the model sets the underlying, technical fundament. As it was mentioned, simuland does not include international factors but the need of their future introduction is extremely vital. Monte Carlo simulation serves the purpose of setting the background and providing a snapshot of the system which could be further extended to include game-theoretic component. With this in mind, future discrete-time simulation introduction is expected.

III. Theoretical Background

This chapter provides theoretical foundations for the conceptual model of the simuland. Selectorate Theory, introduced by Bruce Bueno de Mesquita and colleagues in *The Logic of Political Survival*, proposes an innovative way of thinking about choices being made by decision-makers of various levels. For the use of this model, it provides parsimony needed to simulate behavior which is social in

nature and proposes one, decisive factor for country's choices: state leader's struggle for political survival.

A. Selectorate Theory Overview

The basic idea proposed by Bueno de Mesquita and colleagues is that state's leader is preoccupied with his own survival in the office. This motivation drives all his political choices and therefore, choices of the state. In order to remain in office, leader needs a support of certain number of people. This amount constitutes a Winning Coalition, W, and varies from regime to regime but has always sufficient size to keep leader in office [1]. W itself is a subset of Selectorate, S, which encompasses all people who have real influence on selection of a leader. Selectorate's size varies similarly, depending on regime type.

By supporting a leader, member of S becomes member of W and receives benefits in the form of private or public goods, provided by a leader [2]. This fact makes him support a leader rather than a challenger. This is the element of crucial importance for this model. The W/S ratio constitutes a loyalty norm which is a probability of being a member of successor's Winning Coalition. If W/S becomes smaller, defection to challenger becomes riskier, as there is small probability of being included into challenger's W. Therefore, leader's priority is to keep W as small as possible, if he wants to remain in office. Otherwise, challenger could propose solution which would be attractive for a member of Selectorate and make him defect and become member of W of challenger. Leader's political survival would be endangered.

The size of W/S ratio is the characteristic of political regime. The biggest value is assumed by presidential democracies, as president needs about half of Selectorate to be chosen. Therefore, W/S is about 0.5 [1, p. 54]. Other regimes considered in this model are proportional democracies, first-past-the-post democracies, juntas, rigged elections autocracies and single-party autocracies. More details concerning W/S size will be discussed later.

As it was mentioned, leader provides benefits for his Winning Coalition in the form of private or public goods. First, he collects taxes from residents and then he distributes them, according to the type of regime. In democracies, where W/S ratio is large, revenues are transformed to public goods, including national security, rule of law or education [1, p. 29]. In autocracies, conditioned by small W, benefits take form of private goods, including rents, favorable tax policies or subsidies [1, p. 29]. This is the crucial factor, which would be reflected in this model by one of the variables.

Applications of Selectorate Theory

Authors of *The Logic of Political Survival* offer preliminary ways of applying their theory. They acknowledge problems with measurement of true sizes of W and S in the real world [3]. In this project, certain analyses were made in order to make Selectorate Theory more realistic and informative. Therefore, abstract measurements were replaced with the real-world data.

This is the natural way of making Selectorate Theory applicable. It has already been recognized by scholars. Sebastiano Lustig offers analysis of political stability in Italy, basing on variables of W and S, fueling them with the real-world data [4]. By doing so, he was able to draw conclusions from W/S ratio variations.

B. Conceptual Model

The conceptual model consists of the following steps:

--First, random mother-state is generated with population as a constituting variable. Then, autonomous region within this state is generated, again in population terms. It is calculated with the mother-state/region ratio, derived from the distribution applicable for both democratic and authoritarian states.

--Second, regime type for mother-state is determined: junta, rigged-elections autocracy, single-party autocracy, presidential democracy, proportional democracy or first-past-the-post democracy. Then, Selectorate size is determined for a mother-state, basing on data distribution for particular regime type. Also, Winning Coalition size is determined, following the same logic.

--Third, GDP value is created for a mother-state, basing on data distribution, different for autocracies and democracies. Then, tax rate is determined, following the same logic. Revenues from taxes and non-taxes are separated. The latter aims to show the abundance of country's resources which does not depend on population size. This includes natural resources, tourism or access to the water.

--Fourth, random value of non-tax resources lost to a region is determined (arbitrarily, 1 to 70%). Tax resources lost to a region are determined by multiplication of total tax resources of mother-state by the ratio which constitutes a separatist region. Therefore, it depends on the population loss.

--Fifth, a new Selectorate and Winning Coalition values are determined, depending on a regime type. Selectorate loss is always associated with population lost to a region, while W size is assumed to be constant in the immediate period after region's demand of independence. New W/S ratio is compared with the old one. This comparison constitutes the heart of the model. Also, the total loss of tax and non-tax resources for a mother-state is determined. The fifth point of conceptual model serves pivotal role of drawing conclusions about simulation results.

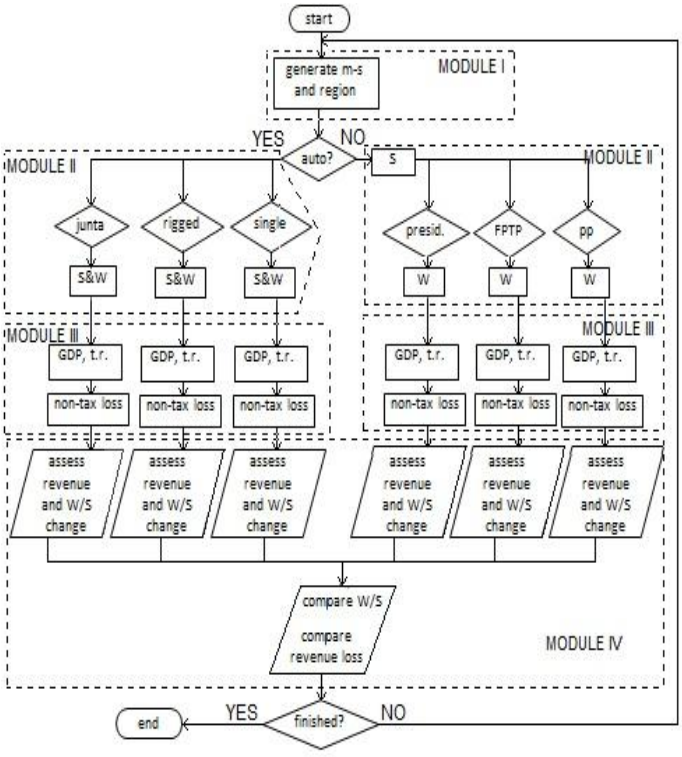


Fig. 1. The Monte Carlo flow chart pictures the conceptual model. "M-s" stands for "mother-state", "auto?" means decision whether a regime is autocratic or not. Names of regimes are abbreviated. "S" stands for Selectorate, "W" for Winning Coalition. "T.r." means "tax rate". Rectangular steps show random generators' processes. Diamonds picture decisions. Parallelograms provide accumulation of simulation-related data.

IV. Operational Model

The operational model contains variables and steps which are being executed in Monte Carlo fashion. Simulation results provide insight about how W/S ratio and resource-endowment would change if region's claim for independence was honored.

A. Variables

Population of Mother-State and Region

The idea behind this variable is to create artificial mother-state and region with independence ambitions. The random generator is based on mechanism derived from real-world data. The total number of 42 mother-states' population values was collected. For all of them, 105 cases of regions demanding independence were identified. As a result, 105 ratios of region/mother-state populations were established. All values were grouped from smallest to largest in order to find the distribution relation. "Cissoid Of Diocles Transform" function was chosen as the best mechanism for random mother-state generator (1). "Holliday" function from "YieldDensity" functions family was used to generate random region/mother-state ratio (2). Resulting random region was produced by multiplying random mother-state with random ratio.

$$y = a((xc - d)^3 \ / (2b - (xc - d)))^{0.5} \quad (1)$$

a = 4.7313E+03; b = -1.2377E+04; c = -5.9954E+02; d = -5.9954E+02

$$y = 1 / (a + bx + cx^2) \quad (2)$$

a = 1.27396E+02; b = -2.1075; c = 8.6848E-03

Regime type, Selectorate and Winning Coalition

Separate simulations were run for each regime type of mother-state. This point is critical in explaining under what regime conditions mother-state leadership is more likely to grant independence to a region.

Juntas

Due to the scarcity of data, only 10 examples of military juntas were considered. Selectorate in this kind of regimes consists of narrow group of colonels or generals and only a fraction of them is in the Winning Coalition [1, p. 71], [5], [6]. Assessing the real size of W is almost impossible for juntas and therefore in this model it would be considered as half of a Selectorate plus one member. Selectorate generator uses sigmoidal "Boltzmann Sigmoid A" function, retrieved from real-world data (3).

$$y = ((a - b)) / (1 + exp((x - c)/d)) + b \quad (3)$$

a = 2.0967E+01; b = 3.34096; c = 6.5440; d = -1.3139

Rigged Elections Autocracies

Just like in the case of juntas, assessing true size of S and W for rigged elections autocracies is a very difficult task. Unfortunately, literature and scholars reach ambiguous conclusions on how to calculate them [7]. In this model, S and W values were assessed basing on "A Selectorate Pilot Study" by Bueno de Mesquita and Smith [8]. The data is very scarce but it constitutes a nucleus for further research. Each rigged elections regime in the dataset was associated with a Selectorate/Population ratio, obtained from "A Selectorate Pilot Study". The ratios distribution follows a trigonometric "Tangent With Offset" function (4). The "Study" further provides W/Population ratios. In order to maintain relation between S and W in this model, resulting W of each state in the dataset was divided by a resulting S. Produced ratios follow "Dose-Response D" distribution from "BioScience" function family (5).

$$y = amplitude * tan(pi(x - center) / width) + Offset \quad (4)$$

amplitude = 9.2611E-02; *center* = 1.0033E+01; *width* = 6.7667; *Offset* = 2.1194E-01

$$y = b + (a - b) / (1 + 10d(c - x)) \quad (5)$$

a = 3.1274E-01; b = 5.1170E-02; c = 2.8781; d = 5.2667E-01

Single Party Autocracies

Assessing Selectorate and Winning Coalition sizes for single party autocracies was probably the most difficult task. As it is pointed out in *The Logic of Political Survival*, S may encompass all dominant party members and W can be a mere fraction of S [1, p. 70]. For this model, it was decided that basing mechanism on the real-world data, no matter how scarce, is better than ambiguous assumptions. Again, tiny dataset constitutes genuine nucleus to be developed. It was assumed that S includes single party members. Party membership/population ratios were established. They follow exponential "Hocket-Sherby" distribution (6). W members are those members of S who have real influence on maintaining leader in office. Since it is nothing more than a lottery, uniformly distributed number was chosen within the range between 13, corresponding to Chinese example, and 200, basing on North Korea's case.

$$y = b - (b - a) * exp(-c(xd)) \quad (6)$$

a = 2.9122E-02; b = 7.2972E-02; c = 3.1605E-02; d = 3.3189

Presidential Democracies

Both S and W are fairly easy to assess in presidential democracies. The former basically includes all voting-eligible population whereas the latter usually consists of 50% of S [1, p. 54]. To be more precise, this model introduces more detailed analysis. To assess a Selectorate size, voting age percentage distribution was retrieved. Each democratic state in the dataset, with Polity Score of 6 and more [9], was assigned with voting age population number. Then, voting age population/population ratios were established. They follow polynomial "3rd Order (Cubic)" distribution (7). This mechanism generates S value as a ratio of total random population. It is the same for all democratic regime types. Winning Coalition size was similarly retrieved from the data which pictures percentage of votes cast for victorious president. They follow "NIST MGH09" distribution function (8) ranging from 38 percent in Nicaragua to 69 percent in Colombia.

$$y = a + bx + cx^2 + dx^3 \quad (7)$$

a = 3.4729E-01; b = 6.7726E-02; c = -4.0093E-03; d = 8.9029E-05

$$y = a\,(x^2 + bx) \,/\, (x^2 + cx + d) \quad (8)$$

a = 5.1038E+01; b = -2.1428E+01; c = -2.0211E+01; d = -1.1415E+01

Proportional Party List Democracies

In this type of democratic regime, Selectorate is established with the same mechanism as used for presidential democracies. Winning Coalition sizes were retrieved from the database of various elections results in party list democracies and correspond to percentage of total votes cast for victorious party which nominated a prime minister as a result. These sizes follow "Simple Equation 19 With Offset" distribution (9).

$$y = a * exp(b/x + cx) + Offset \quad (9)$$

a = 1.2728E+01; b = -1.1065E+01; c = 2.7617E-02; *Offset* = 1.9999E+01

First-Past-The-Post Democracies

Once again, in first-past-the-post regime type Selectorate is generated with mechanism used for presidential and proportional democracies. Here, Winning Coalition size is smaller than in presidential democracies but larger than in proportional democracies. The use of real-world data shows correspondence with the logic set by the theory but fixes its vagueness. Percent of votes cast for victorious party ranges from 26 in India to 53 in Mongolia and follows "NIST MGH09" distribution (10).

$$y = a\,(x^2 + bx) \,/\, (x^2 + cx + d) \quad (10)$$

a = 3.7124E+01; b = -2.5033E+01; c = -2.2736E+01; d = -1.4576E+01

GDP and Tax Rate Variables

The dataset is divided to democratic and authoritarian regimes. For each state within each category GDP values were gathered, together with tax rates. A tax rate multiplied by GDP is a rough estimation of state's revenue from taxed population. It will therefore vary depending on changes in population size after region's declaration of independence. The remaining portion of GDP represents revenues gathered from broadly defined, inherent state's characteristics, not connected to population size, for example natural resources. For democratic states, GDP data follows bio-scientific "Michaelis-Menten Double" function (11), whereas tax rate follows trigonometric "Tangent With Offset" function (12). For authoritarian regimes, GDP data follows "Cissoid Of Diocles Transform" function (13), whereas tax rate follows bio-scientific "Hyperbolic G" function (14). The mechanism generates GDP and tax rate separately, as there are recent scientific proofs that state's economic growth doesn't depend on value of tax rate [10]. This helps to rule out potential correlation.

$$y = ax \,/\, (b + x) + cx \,/\, (d + x) \quad (11)$$

a = -1.2657E+11; b = -7.8692E+01; c = 1.6740E+14; d = 1.4028E+05

$$y = amplitude * tan(pi * (x - center) \,/\, width) + Offset \quad (12)$$

amplitude = 2.5258; *center* = 3.7995E+01; *width* = 8.3572E+01; *Offset* = 1.68098E+01

$$y = a((xc - d)^3 \,/\, (2b - (xc - d)))^{0.5} \quad (13)$$

a = 5.2257E+05; b = -1.8301E+05; c = -1.4625E+04; d = -1.4625E+04

$$y = ax \,/\, (b + x) + cx \,/\, (d + x) \quad (14)$$

a = -7.6514E-01; b = -2.7021E+01; c = 2.1054E+01; d = 9.6871

B. Operational Model

In the Monte Carlo simulation of this model all of its modules and variables are executed simultaneously. The ultimate goal is to compare a new W/S ratio to initial one as well as a total loss of state's GDP resources. This change and loss are potential and represent a situation where mother-state decides to grant independence to a region which declares it. Comparing a potential change and loss among regime types is an ultimate action by which one can answer the question under which regime conditions leaders are more prone to grant independence to a region.

The execution process follows conceptual model's logic and uses mechanisms/distributions described in "Variables" chapter. It consists of 4 modules.

Module I: Random Mother-State and Region Generator

In this module, a random mother-state is generated in terms of population. Similarly, random region/mother-state ratio is generated. Both values are multiplied and random region is created, in terms of population. The mechanism is the same for both democratic and authoritarian regimes.

Module II: Selectorate and Winning Coalition Generator

In this module, separate simulations are run for each of the regime-types. In juntas Selectorate is generated directly by the distribution function. For rigged elections and single party autocracies the ratio of S/population is randomly generated and then multiplied by population of a mother-state. Each of them has its own ratio generator. For all three types of democracy the same S generator is used. The ratio of S/population is generated and then multiplied by a population. Winning Coalition size is randomly generated, following rules set in "Variables" chapter, specific for each type of regime.

Module III: GDP and Tax Rate Generator

In this module, GDP value and tax rate value are generated randomly, following distribution functions different for autocracies and democracies as two branches of regimes. GDP values are multiplied by tax rates values and new "tax-revenue" variables are created. They are then subtracted from total GDP values and new "non-tax revenue" variables are created.

Module IV: W/S Change and Revenue Loss Determination

This is the ultimate module which completes, unifies and summarizes simulations of the first three modules. This module uses the same mechanisms for all regime types.

--First, it multiplies S of a mother-state by region

population/mother-state population ratio. This product is subtracted from the original S size of a mother-state. Therefore, it shows a potential loss of Selectorate if region's independence claims were recognized. Winning Coalition Size is, by definition, assumed to remain constant.

--Second, it multiplies a tax-revenue variable by region population/mother-state population ratio. This product represents total tax resources lost to potentially independent region. For non-tax revenue variable, the module multiplies it by uniformly distributed value within the range between 1 and 70 percent. It pictures percent of non-tax resources lost to potentially independent region. This percent loss is established arbitrarily and intuitively.

--Third, it compares initial W/S ratio to a new one. Then, it sums total loss of tax and non-tax resources and presents it as percentage loss of total GDP resources to potentially independent region. These two products are then used to draw conclusions about simulation results.

V. Results of the Simulation

The table below presents results of the simulation. Here, the percentage change of W/S ratio is assessed. The bigger is this change, the less prone a mother-state is to grant independence to a region. In case of each regime type, 100 iterations were run and average mean was calculated. Tax and non-tax resources loss was calculated in the same fashion. It equaled 31% for autocracies and 29.7% for democracies.

TABLE I
W/S Change Simulation Results

	Regime type	Change in W/S ratio
AUTO	juntas	11.590
DEMO	presidential	10.280
DEMO	FPTP	9.384
DEMO	proportional	9.118
AUTO	single	7.377
AUTO	rigged	7.199

According to the results, juntas are the least prone to grant independence to their regions, while rigged-election autocracies are the most willing. From among democracies, presidential type is most skeptical while proportional type is the most optimistic. Differences between proportional and FPTP democracies, as well as between single and rigged-election autocracies are slight. Critical assessment of these results and resources loss account is provided in the next chapter.

VI. V&V Account

A. Verification of the Model

Due to significance of the role played in this model by random generators, their verification is essential. Here, the case of Winning Coalition generator in proportional democracies is used as a proposal for further developments, preferably when mechanisms are improved with more comprehensive data. The figure below introduces sensitivity analysis, assessing function's reaction to changes in parameters.

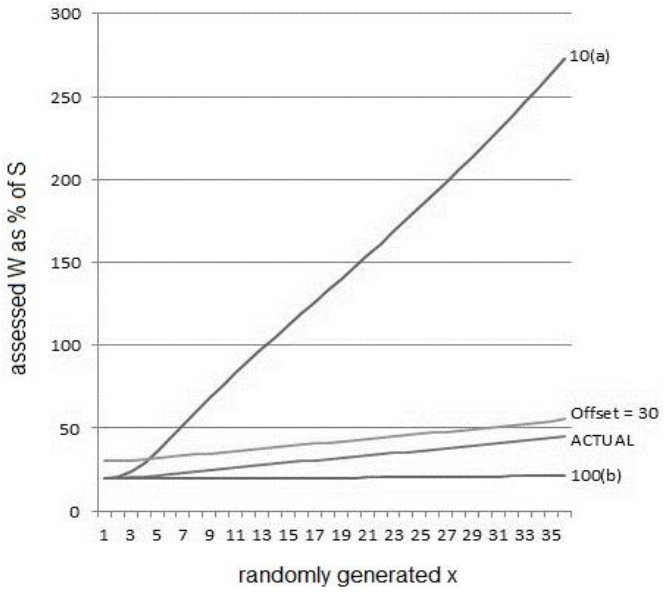


Fig. 2. Sensitivity of proportional democracy W generator-function (9). As it is shown, relatively small increase of *a* brings about extreme results. Similarly, increase in *b* makes function non-predictive and flat. Increase in *Offset* value shifts function in parallel manner. X-axis consists of 1-36 range of x inputs, corresponding to real world dataset of 36 elections cases in proportional democracies.

The W generator for proportional democracies is verified for the given dataset. Without extensive data build-up and research it is impossible to assess to what phenomena are these parameters related. Theoretically, increase in *a* would be observable in presidential democracies, increase in *b* could mean very similar institutional conditions, maybe among states with shared cultural and legal histories, while increase in *Offset* might be observable in FPTP democracies case.

Further data build-up should improve model's validation and overall credibility. What is more, collection of the real-world data would enhance more extensive verification procedures.

B. Validation and Limitations of the Model

The validation of the conceptual model is a comprehensive process. It is fairly easy to be conducted in parts based on Selectorate Theory. There are, however, certain modifications which will have to be thoroughly examined by further developers. The assumption that W remains the same after declaration of independence is intuitively correct but will need a support of the real-world data. But in general, the idea of various implications of Selectorate Theory has already been used, for example to draw conclusions about differences between autocratic and democratic interventionism [11].

Similarly, all random generators constitute nucleuses but have to be reinforced with more comprehensive databases. At the present stage of development, Module III is just a proposal. It is nevertheless the crucial element of any model basing on Selectorate Theory. Future developments must therefore focus on assessing a true conjoined effect of loss in state's revenues and W/S change.

The idea of different reactions of democratic and autocratic leaders towards a region declaring independence is quite hard to validate. There are practically no cases of successful

separatisms with a full consent of a mother-state. Emergence of new states is usually heavily influenced by external processes, like decolonization or wars. Adding international component must therefore be a pivotal goal for further developers.

At this stage, validation may include considering federalism and existence of autonomous provinces as "harbingers" of mother-state's willingness to grant independence to a region in the future. Looking at this phenomenon, from among 6 rigged-elections regimes, 4 are federations [12]. Next, 1 out of 5 single-party dictatorships, China, has autonomous provinces. From among 7 proportional democracies only Austria is a federation but some have forms of self-governing territories. In first-past-the-post democracies India, Great Britain and Canada are significant examples of empowering regions with autonomous capabilities. Then, out of 19 presidential republics only 3 are federal and 2 have autonomous regions. Finally, the junta case in Burma possesses some self-governing territories but it is intuitively not an overall tendency.

These results in general prove that conceptual model catches some part of phenomenon but further research is essential. The serious example of Russia, rigged-elections autocracy, and its attitude towards claims for independence has to be given primary attention. It is obvious that human rights and rule of law factors should be included by further developers and democracies should be even more strictly separated from autocracies. Faroe Islands and Greenland, within democratic Denmark, are strong cases of successful consent for independence. Denmark is proportional democracy and therefore it follows the results revealed by simulation.

Summing up, significant part of validation can be implemented but the rest of it should be taken over by further developers, along with further improvements of the model. Model's fidelity is by now satisfactory.

REFERENCES

[1] B. Bueno de Mesquita, A. Smith, R. M. Siverson, and J. D. Morrow, *The Logic of Political Survival.* Cambridge, MA: The MIT Press, 2003, p. 51.

[2] J. D. Morrow, B. Bueno de Mesquita, R. M. Siverson, and A. Smith, "Selection institutions and war aims," *Econ. Gov.*, vol. 7, 2006, p. 31.

[3] B. Bueno de Mesquita, J. D. Morrow, R. M. Siverson, M. Randolph, and A. Smith, "Testing novel implications from the Selectorate Theory of war," *World Politics*, vol. 56, Apr. 2004, pp. 374-376.

[4] S. Lustig, "Selectorate Theory: solving Italy's instability," *Journal of Political Inquiry*, vol. 5, no. 5, 2012.

[5] "Government and political conditions: background notes on countries of the world: Burma," Political Science Complete, Jul. 2009.

[6] J. Teorell, *Determinants of Democratization: Explaining Regime Change in the World, 1972-2006.* Cambridge: Cambridge University Press, 2010, p. 119.

[7] P. F. Diehl (ed.), *The Scourge of War: New Extensions on an Old Problem.* Ann Arbor, MI: University of Michigan Press, 2004, ch. 2, p. 101.

[8] B. Bueno de Mesquita, and A. Smith, "Dimensions of governance: a Selectorate pilot study," unpublished.

[9] "Polity IV Project." Available: http://www.systemicpeace.org/polity/polity4.htm.

[10] T. L. Hungerford, "Taxes and the economy: an economic analysis of the top tax rates since 1945," CRS Report for Congress, Sep. 14, 2012.

[11] B. Bueno de Mesquita, and G. W. Downs, "Intervention and democracy," *International Organization*, vol. 60, issue 3, 2006.

[12] " CIA World Factbook". Available: https://www.cia.gov/library/publications/the-world-factbook/.

Modeling and analysis of continuous longitudinal data using antedependence models

Sirisha L. Mushti and N. Rao Chaganty
Department of Mathematics and Statistics, Old Dominion University,
Norfolk, Virginia, USA
April 11, 2013

1 Introduction

Longitudinal data, data obtained on same subject at several different time points, are an increasingly common feature in biomedical or clinical trials. The analysis in longitudinal studies usually focuses on how the data change over time and how they are associated with certain risk factors or covariates. Analysis of such data is complex since the responses from the subject observed at t time points, $Y_i = (y_{i1}, y_{i2}, \ldots, y_{it})'$ $i = 1, 2, \ldots, n$, are dependent. These dependencies can be accommodated by choosing an appropriate correlation structures. One such choice of dependency structure is antedependence models.

Definition 1. *Index-ordered random variables $Y_1, Y_2, \ldots, Y_t$ are said to be antedependent of order p, or AD(p) for $0 \leq p \leq t-1$, if Y_k, given at least p immediately preceding variables, is independent of all further preceding variables for $k = 1, 2, \ldots, t$.*

A first-order antedependence random variables $\{\zeta_s : s \in (0, 1, 2, ...)\}$ can be generated by the non-stationary time-series model

$$Y_s = \rho_{s-1} Y_{s-1} + \epsilon_s \tag{1}$$

where ϵ_s are uncorrelated $(0, \sigma^2)$ random variables and $\rho = (\rho_{1}, \rho_{2}, \ldots, \rho_{t\text{-}1})$ are unrestricted autoregressive coefficients.The correlation function is given by

$$\text{Corr}(\zeta_s, \zeta_{s+k}) = \rho_{s}\rho_{s+1}\rho_{s+2}\cdots\rho_{s+k\text{-}1} \qquad \text{for } k \geq 1 \tag{2}$$

Let $X_i = (x_{i1}, x_{i2}, \ldots, x_{it})'$ be the covariates corresponding to Y_i for $i = 1, 2, \ldots, n$. We study the relationship between response Y_i and covariates X_i using the model $E(Y_i) = \mu_i = X_i\beta$ and $Cov(Y_i) = \phi R(\rho)$ where β is the regression coefficients, ϕ is the residual variance of the responses and $R(\rho)$ is the correlation matrix constructed using 2. Standard procedures such as maximum likelihood method gives the closed form expressions for β and ϕ. However, there are complexities involved in estimating the correlation parameters ρ. Hence, we suggested an alternate estimation method for estimating ρ known as Quasi-Least Squares proposed as in Chaganty (1997) and Chaganty and Shults (1999).

2 Estimation Methods

The estimators of β and ϕ are

$$\begin{aligned} \widehat{\beta} &= \left(\sum_{i=1}^{n} X_i' R^{-1}(\rho) X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' R^{-1}(\rho) Y_i\right), \\ \widehat{\phi} &= \frac{1}{nt} \sum_{i=1}^{n} \left(Y_i - X_i\widehat{\beta}\right)' R^{-1}(\rho) \left(Y_i - X_i\widehat{\beta}\right) \end{aligned} \tag{3}$$

We estimate ρ using the two-step quasi-least squares method defined as follows. Consider quasi-log-likelihood function $Q(\theta) = \sum_{i=1}^{n} (Y_i - X_i\beta)' R^{-1}(\rho) (Y_i - X_i\beta)$. We minimize $Q(\theta)$ with respect to ρ to obtain $Step$-1 estimates $\widetilde{\rho}$. Since $Step-1$ estimate $\widetilde{\rho}$ are biased, we solve $\text{tr}\left(\dfrac{\partial R^{-1}(\widetilde{\rho})}{\partial \rho} R(\rho)\right) = 0$ to obtain $Step-2$ unbiased estimates of ρ, say $\widehat{\rho}$. Thus, considering $R(\rho)$ as the antedependence correlation structure, closed form expressions of $Step-2$ estimates of $\rho = (\rho_1, \rho_2, \ldots, \rho_{t\text{-}1})$ are given as

$$\widehat{\rho_j} = \frac{2\sum_{i=1}^{n} z_{ij} z_{i,j+1}}{\sum_{i=1}^{n} \left(z_{ij}^2 + z_{i,j+1}^2\right)} \qquad \text{for} \quad j = 1, 2, \ldots, t-1. \tag{4}$$

3 Results and Conclusions

The estimates obtained using quasi-least squares method can be evaluated asymptotically based on the unbiased estimation theory. In addition, we have evaluated these estimators in small-sample case using the simulations. The quasi-least squares estimators are compared to the standard maximum likelihood estimators using the relative efficiency, ratio of the mean squared error of the two estimators. From the efficiency graphs we have observed that the quasi-least squares method is a good competitor for the maximum likelihood estimators.

References

Chaganty, N. R. (1997). An Alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference*. 63, 39-54.

Chaganty, N. R. and Shults J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference*. 76, 145-161.

Shults J. and Chaganty, N. R. (1999). Analysis of Serially Correlated Data Using Quasi-Least Squares. *Biometrics*. 54, 1622-1630.

DESCRIBING THE ROLE OF CALIBRATION IN SYSTEM DYNAMICS MODELS: SOME ISSUES AND ANSWERS

Ange-Lionel Toba, Old Dominion University- VMASC, Norfolk, VA

***Abstract*—Calibration tools applied to simulation models may have a critical impact on model validity .These tools, when related to System Dynamics (SD) simulations, have a key importance as it may allow the model to mimic real-world systems. This tool helps injecting more confidence into decision making processes. The quality of the model is thus improved given its ability to produce accurate values. This paper explores directions in the use of calibration methods in SD models. In general, a calibration method explores agreements between simulation results and historical data, and then it produces adjustments and evaluations that assess th impacts of these adjustments on these agreements. This paper focuses on describing the calibration process using the VENSIM built-in Optimizer feature.**

***Keywords* — System Dynamics, Calibration, Sensitivity Analysis, Validation.**

I. INTRODUCTION

The process of model calibration is defined by the parameters' estimation as well as their specification. The estimation process appears as trials/experiments seeking those parameters' values to be as close as possible to the real values or, in other words, to be within acceptable tolerance of error (gap between simulated and observed data) (Sorooshian and Gupta 1995). Calibration consists then in finding the values of certain elements of a model, which would replicate the observations from the real system. Thus, the variables and constants in a model, as well as their relationships, require close attention and carefulness in their determination. In this sense, (Oliva 2003) emphasized the absolute necessity of a thorough model calibration for adequate results interpretation. She pointed out to the importance of accuracy in linking the structure of the model to its behavior and the robustness of the model through accurate calibration. The author analyzes the calibration issues encountered at the theoretical, methodological, and technical level and proposes a set of techniques testing the validity of eventual hypothesis made by the modeler. For her, the process of calibration is not only the first step in testing the model, but also a key one in the construction of the model. (Kong, McMahon et al. 2010) also stress the importance of the calibration, especially automatic calibration, of a model in the domain of medicine. They make use of two engineering-based calibration methods: Genetic Algorithm and Simulated Annealing, to find a good goodness-of-fit score for their model LCPM (Lung Cancer Policy Model). This method grants speed to the overall process of calibration and limits the human intervention in the process. This responds to the issue posed by

the constant change of the LCPM, as this automatic calibration would minimize potential bias created by human input. The importance of calibration is also addressed by (Popova and Kercheva 2005) in both their CERES-maize and CERES-wheat models. Their proposed calibration method proved precise enough, as it allowed acceptable validation and accurate forecasting over a consistent time period.

The central objective of this research is to describe a series of steps to follow in order to perform an adequate calibration process, having provided the key variables, most influential in the behavior of the model. The next section will provide a few examples of other suggested calibration methods. The elaboration of the details of our suggested calibration steps and its characteristics will be described in the following section. Within that section, there will be a clear description of the process and assumptions made along the way. The conclusion will compose the last section of the paper.

II. RELATED WORK

The issue of validation through model calibration has been and is still being widely addressed, across various modeling disciplines. (Park and Schneeberger 2003) suggested a procedure for the calibration and validation of their microscopic simulation model in the domain of transportation. The technique was applied using real-world traffic data from Route 50 on Lee Jackson Highway in Fairfax, Virginia, through nine steps. The software used was VISSIM, which in addition provided valuable and significant visualization insight necessary for the validation of their model. Results proved to be satisfactory. In the domain of re-construction, (Parvan, Rahmandad et al. 2012) created a model to assess the behavior of a multitude of construction projects. They examined the consequences of errors due to design on the quality of the constructions. They suggested that a calibration method allows the specification of key variables in the model as well as their parameterization. The objective of this calibration was to minimize errors and provide a realistic representation of the construction projects. (Azcarate, Mallor et al. 2012) present a model trying to replicate the bed occupancy level (BOL) in an intensive care unit. They proposed a method determining the model's parameters which would provide a match with the real system. The calibration problem is expressed as a nonlinear stochastic optimization problem, with the objective of minimizing the output difference between the two systems. An additional objective function, minimum medical intervention, is later used in the algorithm in order to improve the calibration process. This method seeks to include discharge decisions made by medical personnel in the calibration and suggest rules in order to model human decision-making.

III. METHOD

1. Motivation/Problem

Modeling and simulation (M&S) has become a quite reliable tool, providing results to guide system design, decision making and operational philosophies in various disciplines. It is thus crucial to define the limitation of the models and ensure that the results outputted are acceptable under all/most situations (Pace 2004). Finding the match between the theoretical aspect of the model and its realistic counterpart proves to be difficult task. The issue of absence of data may affect the confidence of a model and ultimately its validation. Calibration assists in providing an answer to this challenge. That is, satisfactory parameters' values help ensure that the model and its functions accurately follow the initially intended objective. The validation process establishes the credibility of the model, showing its ability to replicate historical behaviors and provide accurate future patterns. This is a necessary process that verifies that the models considered, the simulation runs and its details are not only correct, mimic the reality, but also reliable.

2. Method

The software employed in this paper to explore the impact of calibrations is VENSIM,

and more specifically, the optimizer embedded in the simulation tool. As an optimizer, after selecting the measurements of interest, the objective is to choose to either minimize or maximize them in accordance with the goal of the simulation. It is important to note that one optimization per variable of interest is possible at a time. The purpose of this test is to match the model with historical data. In doing so, the parameters' values of the model are estimated and the model has achieved the test of reproducibility in the calibration process. It increases the level of acceptability of the model, as well as its confidence (Schade and Krail 2006). The execution of this task is not done without the selection of parameters to assess.

The historical data are entered into the optimizer, as well as all the constants in the model, either in the same sub-model or not. The constants chosen are subject to range constraints. Those constraints can be deduced either through the analysis of the parameters' formulation relationships (conditions imposed by the equations used in the model), studies (past studies which apply to the same case of the model), or face value (expert opinions). The variable of interest is optimized and its behavior is compared to the behavior of the old data. The solver provides estimation for the constant allowing the error (gap between the two curves) to be at its minimum.

The next phase is the sensitivity analysis. This step has the objective of specifying the most influential constants and drawing special attention from the simulationist or decision maker. This test is also performed using VENSIM and its built-in feature. Similar to the previous step, a single variable of interest is chosen, and its behavioral changes are dictated by the selected constants. This test determines which variable is affected should a decision be made, and to what extent is the variation, if any.

This distinction brings more simplicity to the calibration process, regarding the number of parameters to estimate. That is, the parameters with less influence can be fixed at their nominal values, leaving only the ones with important influence on the model's behavior to be examined. This contributes to the reduction of parameter search and ultimately reduces the time to perform the calibration process. This serves the purpose of relevance in the resource allocation, from the user standpoint (Homer and Hirsch 2006) or insight to explain some behaviors, from the model standpoint.

The last step regards the impact of the parameterized variables on the model simulation in the future. This phase takes root in the previous step, as it provides a range of alternatives for the output. In other words, as ranges of parameters have been established and the influence of the different factors determined, the model can generate various output. This gives the possibility of diversifying alternatives to the user. Depending on the horizon of the future to be examined, the model can be run in continuation of the historical data previously found. As for an example, had the past data been run from 1950 to 2013, the length of the time simulation would be extended until, say, 2050 (duration of the simulation run). The range of parameters previously set provides an output range as well, from year 2013 on, showing the window within which the model behavioral pattern is constrained. This output helps in estimating the boundaries of the system. Decisions can be made in an educated manner and with greater efficiency.

Depending on the intended purpose of the model, performing a stepwise calibration may contribute to increase the credibility of the model and its confidence in producing quite relevant simulation results. Though the ability to match/replicate past data inspires some level of trust, the historical fit test alone is not enough of a satisfactory criterion for the validation of the model. In addition, the model has to behave the way it behaves for the right reasons. Its structure should show consistency between the equations' formulations and the output delivered.

The confidence of the model resides, only partially in the reproducibility of past data. The robustness of the model also constitutes an important portion in the confidence building, as the model should deal with extreme conditions (inside and outside of the historical boundaries), expressed by a wider range of circumstances (Sterman 2000). In this sense, the test of sensitivity analysis deals with this issue. The ranges created during the sensitivity test contribute to test the model and expose it to

different scenarios. In terms of policy making, analysts have access to reliable information to make informed decisions. Depending on the initial purpose of the model and the real system it is trying to mimic, the means of designing quality policy eventually rely on the performance of the model simulation. In this case, the accuracy of the estimations of the parameters obtained from the suggested method, the most influential, contribute to the understanding of the problem trying to be solved.

The application of these steps assists in the process of gaining in credibility and confidence. The ability of the model to match/replicate past data logically mirrors its ability to produce accurate simulation results.

IV. CONCLUSION

The barriers in performing model calibration are diverse and important. The confidence of a model acquired after calibration is defined by its ability to generate seemingly real behavioral observation. Failure to achieve this task results in a rejection of the model, which may be deemed as non-valid. The ever presence of uncertainty brings more complexity to the process, given the non-predictability of the real system and the impossibility of the achievement of an error-free model. However, this issue could be avoided, assuming the errors (gap between model and observation) come exclusively from the model calibration (Beven 2006). With that assumption, the method described above appears to be a good prospect for the calibration of our model and others that are similar.

Although the method described reinforces the reliability and validity of the model and ultimately provides managerial guidance, the process of calibration is limited. The acceptable level of tolerance included in the historical fit may be object of subjectivity. The error deemed to be acceptable may vary from one analyst to another, which could create an over-reliance on the subject matter expert (SME). From that point, the validation becomes an exercise based on the sole perception of the simulationist who is supposed to quantify the confidence with which he believes the model may mimic. The step of the sensitivity analysis is also not without criticism, considering the manner it is performed. One important limitation is the fact that only one parameter at a time can be changed, while the other ones remain unchanged. In this sense, the reliability of the variation of the different terms depends on one another. Thus, there are limitations in taking into account the combined effects of variation of more than two elements. This could jeopardize the conclusions drawn from the simulation as some parameters may be wrongly estimated.

These points demonstrate that it is quite unlikely to label any calibration method as the "best". Given the diversity of the modeling purposes and conceptualization, a good calibration is the one that closest align to the analyst purposes.

REFERENCES

Azcarate, C., F. Mallor, et al. (2012). Calibration of a decision-making process in a simulation model by a bicriteria optimization problem. Simulation Conference (WSC), Proceedings of the 2012 Winter, Pamplona, Spain

Beven, K. (2006). "A manifesto for the equifinality thesis." Journal of Hydrology **320**(1–2): 18-36.

Homer, J. and G. Hirsch (2006). "System Dynamics modeling for public health: Background and opportunities." American Journal of Public Health **96**(3): 452-458.

Kong, C. Y., P. M. McMahon, et al. (2010). "Calibration of disease simulation model using an engineering approach." Value Health **13**(1): 157.

Oliva, R. (2003). "Model calibration as a testing strategy for System Dynamics models." European Journal of Operational Research **151**(3): 552-568.

Pace, D. K. (2004). "Modeling and simulation verification and validation challenges." Johns Hopkins APL Technical Digest **25**(2): 163-172.

Park, B. and J. Schneeberger (2003). "Microscopic simulation model calibration and validation: Case study of VISSIM simulation model for a coordinated actuated signal system." Transportation Research Record: Journal of the Transportation Research Board **1856**(-1): 185-192.

Parvan, K., H. Rahmandad, et al. (2012). Estimating the impact factor of undiscovered design errors on construction quality 30th International Conference of the System Dynamics Society, St. Gallen, Switzerland

Popova, Z. and M. Kercheva (2005). "CERES model application for increasing preparedness to climate variability in agricultural planning – calibration and validation test." Physics and Chemistry of the Earth, Parts A/B/C **30**(1–3): 125-133.

Schade, W. and M. Krail (2006). Modeling and calibration of large scale system dynamics models: The case of the ASTRA model. 24th International Conference of the System Dynamics Society, Nijmegen, The Netherlands.

Sorooshian, S. and V. Gupta (1995). Model calibration. Computer Models of Watershed Hydrology. V. P. Singh. Highlands Ranch, CO, Water Resources Publications**:** 23-63.

Sterman, J. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World, McGraw-Hill/Irwin

An Analysis of Various GPU Implementations of Saint-Venant Shallow Water Equations

Joseph C. Miller, III

***Abstract*—The purpose of this project is to analyze the efficiency of various GPU implementations of shallow water equations. Visualization of the Saint-Venant shallow water equations in real-time is computation intensive; sequential single-CPU implementations of the equations yields very poor results of less than a single iteration per second. It is therefore necessary to either perform the calculations in parallel on multiple CPUs or on the GPU. The focus of this paper is the analysis of two methods of GPU calculation: CUDA and HLSL compute shaders.**

***Index Terms*—GPU, CUDA, HLSL compute shaders, Saint-Venant shallow water equations.**

I. Introduction

Real-time visualization of three-dimensional problems is a helpful tool in the Modeling and Simulation environment. Disaster simulations that emulate various natural disasters such as tsunamis, earthquakes and hurricanes make excellent use of real-time visualization to allow the user to see the effects of said disasters in real time. However, visualizing such disasters in real-time is complex and involves rapidly solving large mathematical equations. For example, rendering tsunamis in real time using the Saint-Venant shallow water equations requires the solution of partial differential equations (PDEs) in each cell of a large grid (e.g. 512×512). Unfortunately, currently available single CPUs are not capable of solving partial differential equations on that scale quickly enough to render the water interactions in real time. Two options exist for real time calculations: performing the work on many CPU cores in parallel and performing the work on the GPU.

II. Problem/Motivation

The basis for this paper arose from problems that occurred in another project involving the simulated generation of tsunamis. The Saint-Venant shallow water equations were utilized to calculate the interactions of the ocean in the event a large displacement occurred, e.g. earthquake displacement. However once the equations were implemented sequentially using a large matrix of water heights the performance was found to be abysmally slow with the average speed of rendering being two frames per second or less. The speed of the simulation was severely limited by the computation speed of the CPU; it simply could not perform calculations quickly enough to maintain an acceptable simulation speed. The decision was made to port the simulation from a CPU implementation to a GPU implementation.

The problem then became immediately apparent: What type of GPU implementation would be best? There were two readily available possibilities: implement the shallow water equations using textures passed to the GPU and computing the partial differential equations using High Level Shader Language (HLSL) compute shaders or to port the matrix directly to the GPU and perform the calculations using the Nvidia's Compute Unified Driver Architecture (CUDA).. The purpose of this paper is to investigate each method of implementation and draw conclusions regarding the computational speed and efficiency.

III. Methods

In this particular case visualizing the ocean using the Saint-Venant equations requires the solution of the following partial differential equation:

$$\frac{\partial H}{\partial t}+\frac{\partial U}{\partial x}+\frac{\partial V}{\partial y}=B \tag{1}$$

This system describes water flow as a conservation law with a source term as shown in the following equation [1]:

$$\begin{cases} h_t+(hu)_x+(hv)_y=0 \\ (hu)_t+(hu^2+0.5gh^2)_x+(huv)_y=-ghB_x \\ (hv)_t+(huv)_x+(hv^2+0.5gh^2)=-ghB_y \end{cases} \tag{2}$$

This particular variant of the Saint-Venant equations above is simultaneously well-balanced and positivity-preserving [2].

The two possible implementations used to calculate the solutions to the Saint-Venant partial differential equations utilize either HLSL compute shaders or the CUDA platform. The medium used to render the water on screen is Microsoft DirectX 11. DirectX was chosen as it is compatible with both platforms; DirectX 11, HLSL and CUDA all are coded in C++. The HLSL compute shader allows for program speedups by performing calculations on the GPU when a CPU cannot perform large scale calculations in a timely manner [3]. In this implementation a texture is passed to the shader. This texture is the size of the water grid being visualized, in this case 512×512, where each pixel of the texture corresponds to a grid cell. However, the data of the texture is somewhat specialized. The data is not the standard RGB color data used in normal textures but rather is the height, u and v components of the

Saint-Venant equations. Once passed to the GPU compute shader new values for the height, u and v of each cell are calculated. These updated values are then used to render the latest iteration of the water.

The CUDA method for calculating the shallow water equations is similar in concept but different in execution. Rather than passing a texture to the GPU for computation the data is passed as an array of values which is then bound to a texture. The computations are then performed using a kernel which is a CUDA process that is executed by many threads in parallel on the GPU [4]. Once the kernel has completed calculations of the latest frame the results are used to visualize the water in a manner identical to that of the HLSL implementation.

IV. Conclusion

Once both implementations of the Saint-Venant shallow water equations are complete, the results will be analyzed by comparing the iterations per second, otherwise known as frames per second, of each implementation. As the number of frames per second is dependent on the speed and efficiency of each method it should be a viable indicator of performance.

Real time visualization of large scale 3D problems is extraordinarily computation intensive. However with the growing demand for such visualization from various sources, e.g. the gaming industry and the modeling and simulation industry, among others, it is necessary to explore the most efficient methods of performing these calculations. This paper explored two methods for solving large problems quickly: HLSL compute shaders and the CUDA programming model.

References

[1] Alexander Kurganov and Doron Levy, "Central-Upwind Schemes for the Saint-Venant System," *Mathematical Modeling and Numerical Analysis,* vol. 36, pp. 397-425, 2002.

[2] Alexander Kurganov and Guergana Petrova, "A Second-Order Well-balanced Positivity Preserving Central-Upwind Scheme for the Saint-Venant System," *Communications in Mathematical Sciences,* vol. 5, pp. 133-160, 2007.

[3] Microsoft Corporation. (2012, 3/12/2013). *Programming Guide for Direct3D 11: Compute Shader Overview*. Available: http://msdn.microsoft.com/en-us/library/windows/desktop/ff476331(v=vs.85).aspx

[4] Greg Ruetsch and Brent Oster, "Getting Started with CUDA," *NVIDIA Corporation,* 2008.

Simulation on the clouds

Hamdi Kavak

***Abstract* — This extended abstract presents a concept on developing and running simulation models on computers and mobile devices using web standard called HTML5 and cloud computation. It is believed that this concept will make simulation models accessible to wider audience.**

***Index Terms*—modeling and simulation, html5, cloud computing**

I. Introduction

There has been an increasing attention in Modeling and Simulation (M&S) in recent years. However, many of the current M&S software development products (M&S tools) are *costly*, requires *high-end hardware* and *expertise*. These are common barriers on making wide-use of M&S. This paper proposes a concept of M&S tool for building and running simulation models. This tool allows for cross-platform execution based on HTML5 standard and cloud computing. By using this concept, hardware and software cost are reduced, as it only requires a browser capability that already exists in most of the computers, tablets and smart phones today. It also reduces the expertise required by providing simple-to-use interface.

II. Motivation

In recent years, there has been an increasing interest in M&S. This interest is not only from engineering and sciences but also from various areas of applications such as military, medical, education, transportation and business. People from these areas are interested in M&S in two aspects. One is development of simulation models. The other one is using existing simulation models.

In simulation model development, most of the M&S tools on the market are expensive, require level of expertise and high-end computers to build simulations and run them. In simulation model execution, there are two common ways: 1) running existing model using an M&S tool, 2) running the executable version of the model without requiring M&S tool. In both ways, similar computer setup is needed.

Regardless if the simulation is M&S tool dependent or independent, it runs using an executable file that can be operating system dependent (i.e. exe file) or cross-platform (i.e. .jar file or Java Applet). However, these options are merely runnable from traditional computers excluding new generation mobile platforms.

Mobile platforms are getting more widespread in recent years. In 2012, one of four computers sold worldwide were tablet computers and one of two cell phones sold worldwide were smartphones [3]. Apart from high unit sales, tablet computers are getting integral part in educational system. As such running simulations from multiple platforms, which include mobile platforms, will make M&S accessible to wider audience and under a different business model. Meaning more sales at a lower price (even free). The downside is that a user will be able to build or run simpler simulation models.

III. Methods

In order to build and run simulation models in mobile devices, this paper proposes a concept of a tool called Cloud Simulation Engine (CloudSE). CloudSE marries two emerging technologies: HTML5 and cloud computing. HTML5 is the latest web browser standard that promises running full applications on a browser [1] and all modern browsers support HTML5 standard as default. Cloud computing is "a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet." [2]. CloudSE uses these two technologies in providing cross-platform user interface and simulation engine capabilities respectively.

Error! Reference source not found. illustrates a concept design of CloudSE. User input to be captured as various formats such as text, speech, sketch and touch. Once user input captured, interface converts it to appropriate format for the cloud computing environment and sends it to cloud engine. Cloud engine makes required calculations and returns response back to HTML5 interface. Following sections give more detail about which requirements should be provided by user interface and cloud computing environment.

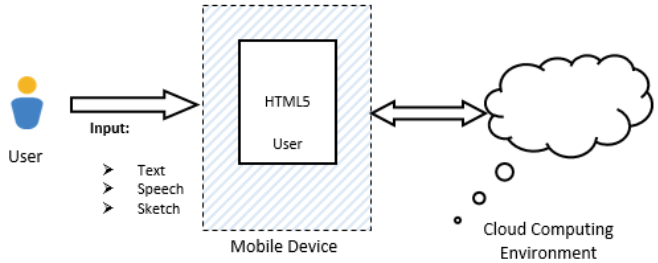


Figure 1: Concept design of CloudSE

Manuscript submitted March 22, 2013.

Hamdi Kavak, a PhD student in Modeling Simulation and Visualization Engineering Department, Old Dominion University Norfolk, VA 23529 USA (hkava001@odu.edu)

A. *User Interface*

User interface part of the tool is based on HTML5 standard. In order to provide HTML5 capabilities, modern web browsers offer JavaScript programming language for computational requirements, HTML and CSS languages for visual interface development. User interface part of CloudSE concept is formed with these three languages to accomplish simulation model development and simulation model execution requirements.

Simulation model development should be accomplished with mostly graphical form of the simulation model elements. User should be able to create, update or delete these elements by various input alternatives shown in **Error! Reference source not found.**. Once the development process is completed, user can save this model on the cloud and can execute anytime.

Simulation model execution can be initiated from user interface after loading the model from cloud environment. User may need to enter simulation model parameters, if asked, then initiate the simulation. Starting from this point cloud simulation environment takes place.

B. *Cloud Simulation Engine*

Cloud simulation engine serves in two ways: simulation model repository, simulation model execution. Simulation model repository keeps models in its database. Simulation engine executes the existing model for each user connected to user interface. If the simulation model requires more computational power, cloud simulation engine creates additional virtual computers dynamically to fulfill this demand and then deletes when execution ends.

IV. CONCLUSION

In this extended abstract, concept design of a cloud simulation engine is presented. Proposed concept and its main components are defined. It is believed that implementation of this concept will boost current simulation industry and create a new business model.

REFERENCES

[1] David, Matthew. HTML5: designing rich Internet applications. Focal Press, 2010.

[2] “Twenty Experts Define Cloud Computing”, SYS-CON Media Inc, http://cloudcomputing.sys-con.com/read/612375_p.htm, 2008

[3] “Market Share Analysis: Mobile Phones, Worldwide, 4Q12 and 2012”, Gartner, Inc., 2013

EXPLORING THE M&S BODY OF KNOWLEDGE

Olcay Sahin,
Modeling, Simulation, and Visualization Engineering
Old Dominion University
Norfolk, Va *23529, USA*
osahi001@odu.edu

***Abstract*— This paper explores the body of knowledge for modeling and simulation (M&S) in order to identify M&S main underline topics. To do so, data from funding organizations and publications is mined and classified base on emerging patterns.**

***Index Terms*— M&S Body of Knowledge, data mining, machine learning.**

I. INTRODUCTION

M&S is applicable to different areas of application, from healthcare to military, areas of study, from social sciences to biology, and purposes, from decision making to training. This variety leads to different ways to build, validate, and use simulations. In addition, this variety, leads to a multitude of M&S topics that seem to overlap with other disciplines such as computer science, system engineering, and software engineering. Ultimately, this variety leads to a lack of agreement of what topics are unique M&S topics or should describe M&S as a discipline.

In order to identify topics from the body of knowledge of M&S, two main data sources are used: 1) funding source, namely National Science Foundation (NSF) and, 2) for M&S publications, namely The Simulation Interoperability Standards Organization (SISO) Simulation Interoperability Workshop (SIW) and the Society for Modeling and Simulation International (SCS) Spring Conference (SpringSim). SIW and SpringSim are used for proof of concept purposes. Other data sources will be explore as part of future work.

II. METHOD

In order to explore the body of knowledge of M&S, a machine-learning based algorithm is used. Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience [1]. This means that a computer learn to identify data patterns after it is trained with existing data.

For training data, papers published in SpringSim, years of 2008, 2010, and 2012 are used. Patterns are identified from SpringSim and SISO SIW using published papers for 2009 and 2011. NSF data has been collected but it is not implemented yet. It is part of the future work.

Table 1 shows the sequence of steps for the machine learning algorithm and tools needed to execute it. Briefly, the algorithm requires to search data from different sources; load that data into a local or many local machines; filter the data eliminating undesired content; categorize the data (papers per year and track); analyze the papers, which implies both training and pattern identification; generate results and visualize results.

Table 1. Machine learning algorithm steps

Steps\Function	Description	Tool
Search	Search within the system	Lucene
	Search on the web	Google Api
Load Data	Select data for store	User Interface
Store Data	Store data in local machine	Local file
	Store data in distributed machine	HDFS [1]
Filter	Filter online data	Alchemy API [2]
	Filter local data	Mahout [3]
Categorize	Categorize online data	Alchemy API
	Categorize local data	Mahout
Analyze	Analyze stored data with using statistical analysis packages. These packages provide: K-means Fuzzy k	Mahout R [4]

	Canopy and dirichlet clustering Correlation Naive bayes Complementary naive bayes . . .	
Generate Results	Generate results in tabular format	User Interface
Visualize Results	Graph results	R

III. RESULTS

Table 2 called confusion matrix which identifies numerically how close/far test data to training data. For instance, out of 37 papers identified for ANSS; 1 is similar to ADS, 6 to itself, 5 to BIS, 10 to CNS, etc. Table 2 shows that papers for ADS, CNS, DEVS, EAIA, HPC, MMS, and SIMAUD have strong similarity with themselves. However that cannot be said about ANSS. ANSS has been found to be similar to CNS and DEVS more than with itself. In other words, people that submit papers to this track do not have clear the purpose of the track.

Table 2. Confusion Matrix

a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**<--Classified as**	
31	1	0	0	0	0	0	3	0	| 35	**a = ADS**
1	**6**	5	10	10	1	2	1	1	| 37	**b = ANSS**
0	0	**3**	1	1	3	1	0	1	| 10	**c = BIS**
0	6	1	**22**	0	1	2	1	0	| 33	**d = CNS**
4	13	4	2	**41**	3	4	2	0	| 73	**e = DEVS**
2	2	1	0	2	**11**	0	2	1	| 21	**f = EAIA**
1	1	0	2	2	0	**32**	0	0	| 38	**g = HPC**
1	0	2	2	2	1	0	**24**	0	| 32	**h = MMS**
0	0	0	0	1	3	1	5	**20**	| 30	**i=SIMAUD**

IV. CONCLUSION AND FUTURE WORK

In this paper, SpringSim and SIW conferences papers are collected and machine learning algorithms have been implemented to proof of concept. In order to find patterns from test data SpringSim papers are trained because, SCS is accepted for being to M&S concept.

For future work, also other M&S papers will be collected and this collection makes the training data more powerful. NSF data has been collected for funding source for the M&S. Findings through funding and publication side of the M&S provides us to which areas are using M&S topics. In this way, lack of agreement can be reduced identifying what topics are unique for M&S.

V. BIBLIOGRAPHY

[1] Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction.Communications of the ACM, 38(11), 54-64.

[2] Orchestr, L. L. C. (2009). Alchemyapi.

[3] Owen, S., Anil, R., Dunning, T., & Friedman, E. (2011). Mahout in action. Manning Publications Co..

[4] Fox, J., & Andersen, R. (2005). Using the R statistical computing environment to teach social statistics courses. Department of Sociology, McMaster University.

Autocorrelation functions applied to the benthic community of the Chesapeake Bay suggest a lack of seasonality

Byron, K.W., Dauer, D.M.

INTRODUCTION

The Chesapeake Bay is the largest estuary in the United States, it covers 11,600 km^2, and the water shed drains 6 states. Population growth in the watershed, overharvesting of fisheries, and development along the shoreline has put the health of the bay in peril (Borja and Dauer 2008; Borja et al. 2010). The Chesapeake Bay Program is the joint effort of state and federal agencies alongside non-profit organizations and academic institutions to oversee the Bay's restoration and protection. A part of protection and restoration efforts, the Benthic Monitoring Program has been in place since 1984 and uses soft-sediment infaunal macrobenthic communities to gauge the health of the Bay. The monitoring program looks at subtidal soft-sediment infaunal macrobenthic communities (hereafter simplified to benthic communities) along with abiotic conditions at the time of sampling. Benthic communities have been shown to be an excellent tool for water quality monitoring (Dauer 1993; Gray 1979) as the organisms respond predictably to disturbance (Pearson and Rosenberg 1978; Rakocinski et al. 2000; Weisberg et al. 1997).

The main products from monitoring programs are biomass, abundance, and richness. These give us a detail picture of the community when sampled (Diaz and Rosenberg 1995; Pearson and Rosenberg 1978; Rakocinski et al. 2000) but help with little more than conceptual ideas of what condition the community is in. A functional metric should be able to provide a stronger understanding of the benthic condition and be readily incorporated into current monitoring and restoration efforts. Secondary productivity is the change in biomass over time of any heterotrophic organism and would is a suitable functional metric of community dynamics (Dolbeth et al. 2012). Brey's empirical model (Brey 1999; Brey 2001) uses depth, temperature, mean weight, and qualitative variables to estimate species-specific productivity to biomass ratios. Here I use Brey's empirical model of secondary productivity to look at the seasonal variation of benthic communities.

METHODS

Data from the Chesapeake Bay Program from 1984 – 1995 were analyzed for temporal patterns in secondary productivity. Sixteen fixed stations in the tidal waters of the Chesapeake Bay within the Virginia state boundaries were selected for analysis. This provided a 10-year time series of uniform data without any missing values. Other stations that were part of this program were added at later dates and stations in Maryland waters were sampled more frequently.

At each station, 3 samples were taken with a box core sampling an area of 225 cm^2 to a depth of 23 cm. Temperature, salinity, depth, and dissolved oxygen were measured concordant with the sediment samples. Sediment cores were wet-sieved through a 0.5mm screen and fixed in bio bags using a 10% formalin solution.

In the lab, organisms were separated from detritus, identified, enumerated, and biomass was determined by ash-free dry-weight.

Secondary productivity was estimated using an empirical model developed for benthic organisms(Brey 1999; Brey 2001). A linear model based on mean individual weight, depth, temperature, and one of three taxonomic classifications, allowed us to estimate a productivity to biomass ratio for each species in a sample. This was multiplied by the total biomass of that species to estimate productivity, which was summed for all the species in the sample. An average of the three samples at each station was used for subsequent analysis.

Utilizing R (Team 2008) a script was developed to test each station for stationarity utilizing a simple regression analysis. Stations showing a positive slope were transformed by differencing, $x_{t+1} - x_t$. Autocorrelation and partial autocorrelation plots were developed for each station and scrutinized for determining terms for the SARIMA model. Confidence intervals for each plot were determined at 1.96*1/sqrt(n) for a cut-off of significant correlation between time lags.

RESULTS

Of the 16 stations analyzed, 3 failed the assumption of stationarity. Of the remaining 13 stations, 10 showed no evidence of statistically significant seasonal patterns and so subsequent analysis was terminated. Of the 3 stations that did show evidence of seasonality, 2 had a single time lag correlation and 1 had a 3-time lag correlation. A biological explanation for these patterns is not readably available and subsequent analysis to understand these trends will be required. The three stations that failed stationarity were subsequently differenced and showed significant correlations at multiple time lags, also requiring a stronger understanding of the biological patterns involved.

DISCUSSION

The majority of the stations analyzed showed no significant temporal trends. This suggests that rates of productivity are relatively constant for the benthic community, or at least vary less than the variation in the benthic community from year to year. The stations that did exhibit some pattern were in an unexpected fashion. Seasonality should be due to temporal fluctuations over time which would smooth into a periodic function with highs in the summer and lows in the winter. This results in an expected negative correlation between t, $t \pm 2$ and a positive correlation between t, $t \pm 4$. Positive correlations at t, $t \pm 1$ suggests secondary productivity is fairly constant for these stations and correlations between t, $t \pm 3$ will require further analysis to fully understand. The overall conclusion from this analysis is that seasonal fluctuations in the productivity in soft-sediment infaunal communities are not likely to have a significant effect on long-term analysis and understanding of the benthic processes.

REFERENCES

Borja, A., and D.M. Dauer. 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators* 8: 331-337.

Borja, A., D.M. Dauer, M. Elliott, and C.A. Simenstad. 2010. Medium- and Long-term Recovery of Estuarine and Coastal Ecosystems: Patterns, Rates and Restoration Effectiveness. *Estuaries and Coasts* 33: 1249-1260.

Brey, T. 1999. A collection of empirical relations for use in ecological modelling. *Naga* 22: 24-28.

Brey, T. 2001. Population dynamics in benthic invertebrates. A virtual handbook. *http://www. awi-bremerhaven. de/Benthic/Ecosystem/FoodWeb/Handbook/main. html. Alfred Wegener Institute for Polar and Marine Research, Germany*.

Dauer, D.M. 1993. Biological Criteria, Environmental-Health and Estuarine Macrobenthic Community Structure. *Marine Pollution Bulletin* 26: 249-257.

Diaz, R.J., and R. Rosenberg. 1995. MARINE BENTHIC HYPOXIA: A REVIEW OF ITS ECOLOGICAL EFFECTS AND THE BEHAVIOURAL RESPONSES OF BENTHIC MACROFAUNA. In *Oceanography and Marine Biology - an Annual Review, Vol 33*, ed. A.D. Ansell, R.N. Gibson and M. Barnes, 245-303. London: U C L Press Ltd.

Dolbeth, M., M. Cusson, R. Sousa, and M.A. Pardal. 2012. Secondary production as a tool for better understanding of aquatic ecosystems. *Canadian Journal of Fisheries and Aquatic Sciences* 69: 1230-1253.

Gray, J.S. 1979. Pollution-Induced Changes in Populations [and Discussion]. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 286: 545-561.

Pearson, T.H., and R. Rosenberg. 1978. MACROBENTHIC SUCCESSION IN RELATION TO ORGANIC ENRICHMENT AND POLLUTION OF THE MARINE ENVIRONMENT. *Oceanography and Marine Biology an Annual Review* 16: 229-311.

Rakocinski, C.F., S.S. Brown, G.R. Gaston, R.W. Heard, W.W. Walker, and J.K. Summers. 2000. Species-abundance-biomass responses by estuarine macrobenthos to sediment chemical contamination. *Journal of Aquatic Ecosystem Stress and Recovery* 7: 201-214.

Team, R.D.C. 2008. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Weisberg, S.B., J.A. Ranasinghe, L.C. Schaffner, R.J. Diaz, D.M. Dauer, and J.B. Frithsen. 1997. An Estuarine Benthic Index of Biotic Integrity (B-IBI) for Chesapeake Bay. *Estuaries* 20: 149-158.

Homeland Security | Military Modeling and Simulation

VMASC Track Chair: Dr. Barry Ezell
MSVE Track Chair: Dr. Bharat Madan

Representing the Ballistic Missile Defense System using Agent-based Modeling
Author(s): Christopher Lynch, Saikou Diallo, and Andreas Tolk

Validation Report of Composer Propagation Simulation: Case Study of Camp Blanding, FL
Author(s): Nicholas Wright, and Robert Kewley

Game Theory and the North Korea Nuclear Puzzle
Author(s): Shiwei Jiang

Invited Paper Submission
Representing the Ballistic Missile Defense System using Agent-based Modeling

Christopher J. Lynch, Saikou Y. Diallo, Andreas Tolk
Old Dominion University, Norfolk, VA, United States
clync008@odu.edu, sdiallo@odu.edu, atolk@odu.edu

KEYWORDS: Agent Based Modeling, Ballistic Missile Defense System, MS-SDF, C2BMC, Missile Defense Agency

ABSTRACT

This paper seeks to explain the applicability and advantage of modeling and simulating the Ballistic Missile Defense System (BMDS) using the Agent-based modeling (ABM) paradigm. This is accomplished through the application of the Modeling and Simulation – System Development Framework (MS-SDF) which provides a framework for capturing the BMDS and creating a model that integrates the human-in-the-loop decision maker into the design. The addition of the human-in-the-loop decision maker is an essential component of the BMDS and is a gap that is has not been sufficiently addressed by other works. ABM allows for the specific elements within the BMDS to be represented and allows for the creation of a highly-configurable simulation environment. The process of creating the model and simulation is provided through the MS-SDF. A reference model is used to identify the components and the characteristics of the BMDS. A conceptual model provides the plan for creating the simulation based on the material contained in the reference model and the simulation is then built based on the conceptual model. This process provides a level of traceability to the model that helps to verify and validate the model and simulation.

1. INTRODUCTION

The BMDS is an integrated system consisting of elements and supporting efforts linked to a command, control, battle management, and communications (C2BMC) network. The purpose of the C2BMC network is to provide operational commanders a link between the sensors and interceptor missiles contained in their areas of responsibilities (AOR). Specifically, the C2BMC network is designed to provide decision making aid to combatant commands (COCOM) by creating an "optimized layered missile defense against all ranges of threat" through the integration and synchronization of missile defense systems [1: p. 8]. This goal is accomplished by processing sensor data from all available sensors (ground-, sea-, and space-based sensors), possibly from multiple military services, and providing a system track that is "accurate enough for targeting and guiding interceptors to lethal warheads" to the decision makers [1: p. 8]. The layered missile defense system provided by the C2BMC allows for more effective tracking on incoming enemy ballistic missiles and provides a wider range of interceptor capabilities for combatant commanders to use in defending against the incoming enemy missiles [1].

A model and simulation of the BMDS was developed using the MS-SDF developed by [2]. The MS-SDF provides a methodology that can be applied to modeling the BMDS by capturing the perspective of stakeholders and subject matter experts (SME), as well as the perspective of the modeler. The stakeholders' and modelers' perspectives are represented by a well-defined set of assumptions and constraints. Insight into the modeling of the BMDS is gained through the use of a reference model. Additional insight and the selection of modeling questions are provided through the use of a conceptual model which provides explicit information on how to take the modeling questions and create the simulation. The MS-SDF provides a measure of traceability between the requirements and the implementation to the modeling process. The clients for this project were from the Missile Defense Agency (MDA) and they served as the SMEs during the design process. They also provided the modeling questions for the conceptual model (discussed in Section 5.1) and they provided many of the assumptions and constraints contained in the reference and conceptual models.

The goal of this paper is to present the design and implementation of a model and simulation of the BMDS with a specific focus on the role of the C2BMC using the MS-SDF. However, the purpose of the contained model and simulation is to provide proof-of-concept that Agent-based modeling (ABM) can be used to provide additional benefits to the modeling and simulation of the BMDS. The model represents the system of ground-, sea-, and space-based elements and supporting efforts that make up the C2BMC network. The simulation is a prototype designed to test and evaluate the effectiveness of different configurations of the BMDS element configurations and capabilities in defending against ballistic missile attacks. ABM is used as the modeling paradigm for this model in order to represent the individual elements, threats, and the C2BMC network that makes up the BMDS.

2. RELATED WORK

Ballistic missile defense has been the topic of many academic evaluations, although not many results are published in journals or conference proceedings due to the often-classified nature of the results of such studies. However, the academic approaches themselves and the topics of research are not always classified, so that a significant body of work had to be evaluated before the proposed concepts were developed.

The Extended Air Defense Simulation (EADSIM) and the Extended Air Defense Test Bed (EADTB) belong to one of the better published system families utilized for ballistic missile defense systems and have been used for studies supporting organization like the Ballistic Missile Defense Organization (BMDO) and the US Army Space and Strategic Defense Command (USASSDC). Both were used in the NATO Active Layered Theater Ballistic Missile Defense (ALTBMD) studies, augmented by other national contributions. The emphasis of these simulations was often the detail simulation of physical processes of tremendous importance for effective missile combat. The mathematical details for missile flight simulation have been composed in a textbook by Strickland [3]. A good overview of recent work is given by Ender et al [4]. The focus shifted from high-detailed analysis of systems and missiles to their interplay in a system of systems coordinated by a common command and control effort. However, the human component was still not sufficiently taken into account.

Niland [5] was among the first to discuss the idea to use agents in support of such tasks within simulated environments. The ideas were generalized and discussed in more detail by Bharathy et al. [6]. The general applicability for agent based modeling approaches in support of ballistic missile defense is described in detail by Garrett et al. [7]. What was still missing was a framework that allowed for integrating the human in the decision making loop as an intelligent software agent. This was the gap formulated by the BMDO that has been primarily addressed by the proposed effort.

3. THE ROLE OF AGENT-BASED MODELING

ABM is a bottom-up modeling paradigm that defines the individuals within a system and then observes the overall system behavior based on the interactions of the individuals and the environment. ABM allows for the BMDS to be represented using four main components: entities, interactions, behaviors, and the environment. Entities include the actors and elements within the BMDS. Interactions are the various relationships between the entities. Behaviors are the actions that can be taken by each individual entity and the environment. The environment is the representation of the real-world system in which the entities reside and the characteristics and attributes that make up this real-world representation.

Garrett et al. identify the usefulness of ABM through its ability to simulate "actions and interactions of autonomous individuals or systems in a shared environment" [7: p. 8]. The ABM modeling paradigm contains advantages for modeling the BMDS environment since the BMDS is an environment of conflict comprised of many different decision-making entities. The agents act to achieve a set of goals based on their individual and incomplete perception of the environment. ABM emphasizes the behavioral relationships of the entities through the established characteristic of each entity. This differs from equation-based approaches which rely on a strict formulation of relationships [7].

Russell and Norvig define an agent as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators" [8: 32]. The use of agents in ABM provides a number of benefits for modeling the BMDS. First, agents are *social***.** This means that agents have the potential to exchange information of any type between each other. Second, agents are *autonomous*, which means that they function based on a defined set of states and behaviors. Russell and Norvig define autonomy as the "extent to which an agent relies on the prior knowledge of its designer rather than on its own percepts" [8: p. 37]. Third, agents are *situated.* They can sense and perceive their environment and use sensors to find out additional information about the environment. The agents can move throughout the environment and they have the capacity to act on objects and other agents. Fourth, agents are *pro-active*. This means that agents have the ability to plan and conduct operations in order to take initiative and not simply respond to their environment. Finally, agents are *flexible*. This means that the agents can adapt to new situations and that they can observe whether their actions are contributing to reaching their goals. This includes the ability of learning from their actions as well as actions from other agents.

4. REFERENCE MODELING

The reference modeling phase of the MS-SDF consists of three steps: exploring the reality/problem situation, creating the reference model, and establishing assumptions. The reference model is defined as "an explicit model of a real or imaginary referent, its attributes, capabilities and relations, as well as governing assumptions and constraints under all relevant perceptions and interpretations" [2]. The primary goal of the reference model is to capture everything that is known and assumed about the problem. Requirements and theories are captured in this step and the model is checked for inconsistencies, over-defined areas, and under-defined areas [2]. The reality/problem situation, the reference model, and the assumptions will be addressed within the following subsections as they are related to the modeling of the BMDS.

4.1. The Reality/Problem Situation

The BMDS is designed to be able to detect, track, engage, and destroy incoming ballistic missiles of all ranges [9]. The BMDS is an "integrated, [and] 'layered' architecture" that is supposed to handle ballistic missiles of any variety of speed, range, size, and performance values. Therefore, the "layered" architecture of the BMDS allows for "multiple opportunities to destroy missiles and their warheads" before they can successfully reach their intended targets [9: p. 1]. The BMDS is comprised of a network of elements (ground-, sea-, and space-based) which include sensors, interceptors, and the C2BMC. All of the BMDS elements, the C2BMC, and the threats are *autonomous* and function based on a specified set of states and behaviors. The elements that are used by the BMDS will be outlined in the following subsection. Then a description of the threats and the environment and their effects on the BMDS are provided in the following subsections.

4.1.1. BMDS Elements

The elements that make up the BMDS are positioned on ground, sea, and space. In reality, there are many different types of elements that do similar tasks. For example, both the Cobra Dane Radar and the Upgraded Early Warning Radar are stationary, ground-based, phased-array, all-weather, and long-range radars that provide midcourse coverage for the BMDS [10, 11]. However, there are still essential differences between the two elements that make them unique, such as the number of faces that the radar has and the diameter of each face [10, 11]. It is not the goal of this model to provide an agent type for each specific type of element within the BMDS, although this is technically possible if so desired by the user. Instead, the current version of the model uses a generalized element type which can have its specific features adjusted based on the specifications of the user.

In addition, an element can consist of a sensor only, a weapon system only, or consist of both a sensor and a weapon system. A weapon system is an element which contains a fire control node and has the capability to launch an interceptor. A sensor is simply a device that is used to track, discriminate, classify, or search for a ballistic missile. The weapon system has the additional characteristics of having a magazine (number of interceptors), a minimum and maximum range, and it has specific types of ballistic missiles that it can intercept (i.e. short-range and medium-range ballistic missiles only). These elements are *situated* since they can sense and act on other agents.

The C2BMC network provides the links between the sensors and the interceptors that allow operational commanders the ability to counter incoming threats [9]. The C2BMC creates the "layered" aspect of the BMDS [12]. A "common, single, integrated ballistic missile picture" is created by the integration of data that flows into the C2BMC from all of the various BMDS elements [12: p. 1]. Information sharing across the BMDS is also a critical function of the C2BMC [12]. This relates to the *social*, *flexible*, and *pro-active* features of agents that were identified in Section 3. Additionally, the following functions of the C2BMC have been identified by [1: p. 9]:

1. Communication and connectivity links between BMD elements;
2. Battle management functions that maximize BMDS effectiveness while minimizing the number of weapons expended;
3. Correlated and optimal system tracks for enemy missiles based on data from multiple defensive sensors;
4. Real-time battle awareness; and
5. Advanced battle planning capabilities enabling BMDS elements to be placed in ideal locations in anticipation of a battle.

4.1.2. BMDS Threats

Threats to the BMDS include short-range, medium-range, intermediate-range, and intercontinental ballistic missiles. The type of ballistic missile is important because it determines the list of targets that the missile is able to reach based on the distance that the missile can travel. Ballistic missiles traverse sequentially through four phases during their flight trajectory [9]. The phases include the boost, ascending, descending, and terminal phases, respectively, shown in Figure 1.

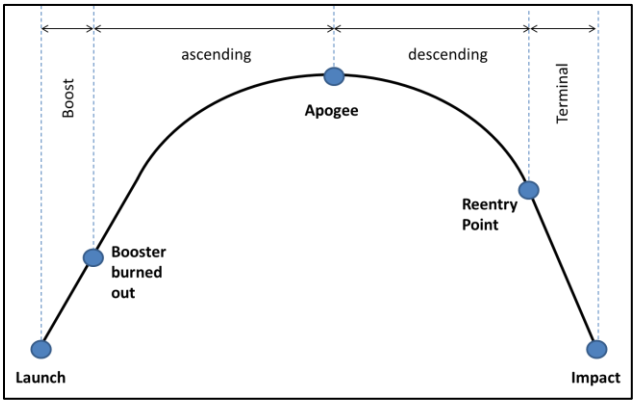


Figure 1: Phases of BMDS

The differences between these phases are important because each of the phases of flight for the ballistic missile means something different for the BMDS. The *boost* phase is the easiest phase to detect and track a ballistic missile due to the light and heat emitted during takeoff. However, this is also the hardest phase in which to engage the missile since the typical length of the boost phase only lasts for about one to five minutes [9]. In addition, missiles are normally launched from terrain that is controlled by opponents so that

an engagement in this early phase is likely to lead to an engagement with hostile forces as well.

The *ascending* phase lasts until the apogee of the flight trajectory is reached. If the ballistic missile can be destroyed before reaching the end of this phase, then the remaining BMDS network resources do not have to be expended in countering the threat [9]. The *descending* phase leads to the reentry point. Many descriptions call this the midcourse phase, which sometimes includes the ascending phase as well. The midcourse phase is the longest phase of the ballistic missile flight path and generally offers "several opportunities to destroy the incoming ballistic missile outside the earth's atmosphere" [9: p. 2]. This offers an additional advantage that any debris from the intercept will burn up in the atmosphere [9]. The *terminal* phase is the final phase of the ballistic missile trajectory. This is the final opportunity to intercept the incoming ballistic missile and it offers little time for an additional interceptor to be launched if the first fails [9]. The interception will now be occurring close to its intended target and ballistic missile debris will fall near this location. However, from a political standpoint, the terminal phase is the least controversial phase for engaging a ballistic missile because it is clearly self-defense from the targeted units.

4.1.3. BMDS Environment

The BMDS environment defines a number of characteristics which relate to specific scenarios that could involve the BMDS. The most important characteristic that is related to the environment is the location of each BMDS element, as well as the starting locations and the targets of each incoming ballistic missile. The time, distance, and speed metrics are set as seconds, meters, and meters per second respectively based on commonly used metrics for missile flight. The location of the threats and elements and the distance between them is a critical component of modeling the BMDS because it determines the total flight time of the incoming ballistic missiles. In other words, this determines the total time that the BMDS has to respond to an enemy threat. The communication capabilities are also established through the environment. BMDS elements can communicate via hard lines (ground-based communication), radio broadcasts, or through communication satellites.

4.2. The Reference Model

The reference model contains everything that is known and assumed about the BMDS. The reference model seeks "completeness" in capturing the BMDS [2: p. 5]. A large portion of the reference model was made up of MDA Fact Sheets that were approved for public release through the MDA website. These fact sheets provided information into the purpose of the BMDS as well as the purpose and functionality of the C2BMC. Additionally, these fact sheets provided information on specific types of ground-, sea-, and space-based elements from which generalizations of these element types were constructed for use in the simulation. The MS-SDF also calls for SMEs to be consulted with during the reference modeling phase. The SMEs that were consulted with during the formation of this model were from the MDA. These SMEs provided additional insights, requirements, and feedback in constructing the reference model.

4.3. The Assumptions

The assumptions relating to the BMDS come from all of the stakeholders involved in the design process, including both the modeler and the SMEs. Two of the assumptions that were identified for the model included ignoring the effects of debris for intercepted ballistic missiles and for both the enemy and friendly forces to disregard the effects of weather on the engagement process. Table 1 provides additional assumptions that are specifically related to the ground-based, sea-based, and space-based elements, as well as the C2BMC and the enemy force. This table shows that the C2BMC has no sensor or weapon system capabilities that are under the direct control of C2BMC and it also shows that the space-based elements are more restricted than the ground-based and sea-based elements. The enemy force is assumed to follow some set of rules of engagement; however, the enemy force rules of engagement are different than the BMDS rules of engagement. It is also assumed that the enemy force does not have any sensors or radar capabilities that it can use to detect the location of BMDS interceptor locations.

Table 1: Assumptions – BMDS Elements and Enemy Force

	Ground-based Element	Sea-based Element	Space-based Element	C2BMC	Enemy Force
Sensor/Radar Capability	X	X	X		
Weapon System Capability	X	X			X
Communicate with Other Elements	X	X	X	X	
Obey the Chain of Command	X	X	X	X	
Obey the Rules of Engagement	X	X		X	X

5. CONCEPTUAL MODELING

The conceptual modeling phase of the MS-SDF consists of three steps: (1) formulating the modeling question, (2) creating the conceptual model, and (3)

establishing constraints. The goal of the conceptual modeling process is to capture the "system's parts and relationships" [2: p. 1]. Modeling questions are designed to be asked of the reference model in order to determine if the necessary entities, properties, rules, and assumptions have been identified in the reference model [2]. If the reference model does not contain the necessary components then the reference model needs to be extended or additional assumptions need to be added [2]. These new assumptions are referred to as "constraints" [2: p. 10]. The goal of the conceptual model is to remain "consistent" with the reference model [2: p. 4]. After the model is deemed consistent with respect to the reference model, the "conceptual model can reflect the paradigm chosen… to answer the modeling question" [2: 10].

5.1. The Modeling Question

The modeling questions to be asked of the BMDS model were provided by the client. These questions formed the overall use case for the model. The main questions that resulted from this use case and that were to be asked of the model included:

1. What element configurations lead to the interception and destruction of all incoming threats within the scenario?
2. What communication configuration capabilities among the elements lead to the destruction of all incoming threats within the scenario?
3. What rules of engagements (ROE) lead to the destruction of all incoming threats within the scenario?

The identification of the modeling questions now allows for the conceptual model to be constructed. Based on the material contained in the reference model, all three of these questions could be addressed. The conceptual model could now be constructed to specifically address the modeling questions listed above.

5.2. Creating the Conceptual Model

The conceptual model captures how the modeling questions will be answered [2]. The Unified Modeling Language (UML) was used to construct the conceptual model. UML captures the conceptual model in the form of classes, behaviors, and structures from the information captured in the reference model. UML provides a standard framework for "visualizing, specifying, constructing, and documenting" the artifacts of a software system [13].The standardized syntax that is provided through the use of UML removes any "ambiguity associated with natural language" and makes the consistency checking process easier to conduct [2: p. 10].

The first stage of creating the conceptual model from the reference model was to describe the purpose of the model through a use case diagram. The purpose of the use case diagram is to "describe the functional requirements" of the BMDS, to provide a "clear and consistent description" of what the BMDS does, and to provide traceability between the functional requirements and the classes and operations in the BMDS [13: p. 58]. The SMEs provided the primary use case for the simulation prototype. The primary use case is shown in Figure 2. This use case shows that the role of the user is to setup a specific scenario and start the simulation and then for the BMDS to capture the engagement process of an incoming ballistic missile from the launch to the interception of the ballistic missile. The user was then to be able to view the results of the simulation based on the starting conditions.

The second stage of creating the conceptual model was to create a sequence diagram that represented the order of operations that were to be taken once an incoming threat was detected. Sequence diagrams display "how objects interact with each other" in a time sequential order [13: p. 174]. These diagrams show the specific actors that are involved in the sequence and messages are used to show the communication between the actors. A sequence diagram has been provided in Figure 3 which shows the engagement process of the BMDS. This diagram shows that for threat that is launched the BMDS sensors try to detect, track, classify, and discriminate the threat. The C2BMC is notified and constantly updated and then the C2BMC assigns a weapon system to engage the threat. The sensors view the result of the engagement and notify the CBMC.

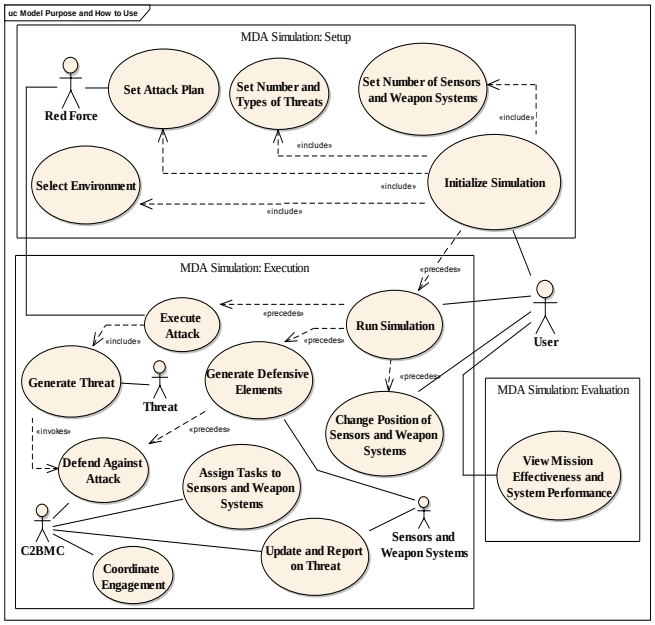


Figure 2: Purpose of the BMDS Model

The final stage of constructing the conceptual model was intended to explicitly capture the functions and attributes that were associated with each of the BMDS

elements. A UML class diagram was used to describe "the static view of [the BMDS] in terms of classes and relationships among the classes" [13: p. 90]. The C2BMC, the threats, and all of the BMDS elements were given their own unique classes containing the attributes and operations associated with each of these elements as described within the reference model. *Attributes* describe the "characteristics of the objects" [13: p. 92] and can also "describe the state of the object" [13: p. 95]. The *operations*, or functions, are used to "manipulate the attributes or perform other actions" [13: p. 95]. An example of some of the attributes that are included for the BMDS elements, the C2BMC, and the threats can be seen in Table 1 which contains several assumptions contained in the reference model. All of the attributes and operations that were defined for each class were compared against the reference model to make sure that no inconsistencies existed between the two models. UML activity diagrams were created to display the actions and states that each element within the model can take.

The agents also have the attributes listed at the beginning of this paper; they are social, autonomous, situated, and pro-active. The BMDS elements and the C2BMC have the ability to communicate with each other during the engagement. Autonomy is created based on the state diagrams of the BMDS elements, the threats, and the C2BMC. The BMDS sensors are situated because they have the ability to detect other objects within the environment. Pro-activeness is best seen in the C2BMC which has the capability to plan engagements and assign the appropriate weapon systems to engage the threats.

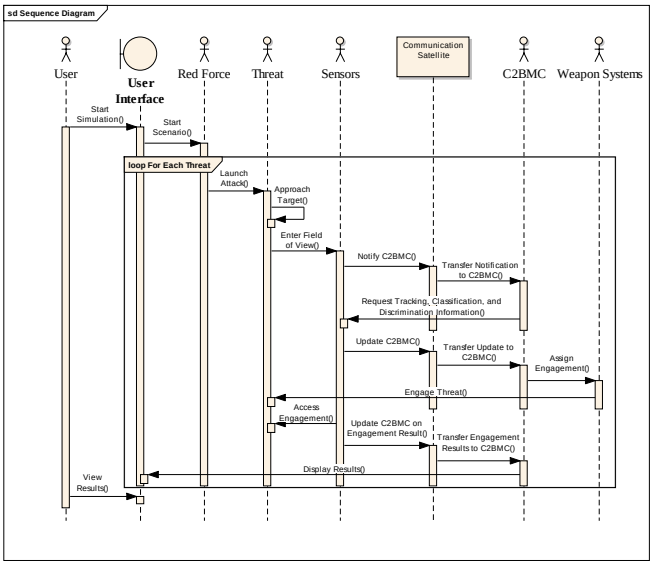


Figure 3: The BMDS Engagement Process

The MS-SDF allows for the conceptual model components to be easily traced back to the original requirements of the model. This allowed the SMEs to verify and validate the conceptual model before the actual simulation was developed. The SMEs validated that the conceptual model was defined well enough to be able to answer the modeling questions that were identified for the model. The SMEs also verified that the conceptual model complied with the requirements and specifications placed on the model during the reference modeling process. During this verification and validation process several additional assumptions were made and they were added to the reference model as constraints. These are addressed in the following section.

5.3. The Constraints

Constraints are new assumptions that are added to the reference model during the conceptual modeling process. The conceptual model is then checked against the reference model with respect to the new assumptions to make sure that conceptual model remains consistent with the reference model [2]. One of the main constraints added to the model deals with the movement of ground-based and sea-based elements. In reality several of these elements, such as the Aegis Ballistic Missile Defense, are mobile; however, a constraint was added that stated that these elements would remain stationary for the duration of the engagement process. Another constraint dealt with the potential set of objects that radar could detect and mistake for a ballistic missile. For example, a friendly plane could appear on radar during an attack and be mistaken for a missile. Therefore, a constraint was added that stated that only ballistic missiles and interceptors would be in the air during the scenario.

A constraint was added to create a generalization of COCOMs and AORs instead of specifically representing a COCOM such as US Strategic Command (USSTRATCOM). Since one of the goals of the model was to test different configurations of BMDS elements and to view the impact of the configuration on the engagement process, a constraint was created that specified that a COCOM and an AOR were defined as the BMDS elements that could communicate with the C2BMC. The C2BMC is assumed to represent the COCOM. It was further assumed that a single element could only have direct communication with a single C2BMC.

6. THE SIMULATION

The simulation was implemented using AnyLogic version 6.7 by XJ Technologies. AnyLogic was selected because it allows for the creation of an agent-based model that defines behavior through the use of state charts. This allows for a smooth transition from the conceptual model (the UML diagrams) to an ABM. The class diagram provided the necessary parameters and functions for creating each of the agents that were being modeled (the BMDS elements, C2BMC, and the threats). The sequence and activity diagrams provided all of the necessary

information for constructing the behavior of each of the agents and creating state charts to represent the current state of the agent and all of the possible actions that the agent could take while in that state. The goals of the agents are clearly reflected in each of their state charts. For example, the goal of the threat is to destroy its target and its state chart represents the movement that it takes in traveling to its target and the current state of flight that the threat is in (boost, ascent, descending, or terminal). AnyLogic also provides a three-dimensional option for constructing the environment. This is a valuable attribute when modeling the BMDS as the ballistic missile flight paths are elliptical and all of the BMDS elements need to be able to interact in three dimensions.

The simulation allows for multiple instances of each of the BMDS elements, the C2BMC, and the threats to be generated for a scenario. The elements can be placed at random or at specified locations defined for the scenario. Each threat is assigned a starting location, a type, and a target at the start of the simulation and they are launched based on the launch method specified for the scenario (launched individually, in small groups, or all are launched at once). The BMDS sensor elements strive to detect, track, classify, and discriminate all threats while constantly updating the C2BMC on the threat positions and velocities. The C2BMC processes the data from the sensor(s) and assigns a BMDS weapon system to engage. At the end of the simulation a number of metrics are output for analysis. These include the starting locations of each element, the number of threats launched, and the number of threats that were successfully intercepted based on the configuration of the BMDS. In addition, specific information on which sensors detected, tracked, classified, or discriminated each threat was also provided along with the time that each of these functions occurred. The BMDS weapon system that intercepted the threat is also provided so that an identification of the strong and weak points of the configuration can be determined. For example, a sensor that provided very little support in detection or tracking could be identified and relocated. Then the scenario can be rerun to test the new configuration.

The execution of the simulation allows the user to follow the BMDS process in defending against incoming ballistic missile attacks. The BMDS elements used in the scenario are placed on a map that displays the AOR(s) being used within that scenario. The user has the ability to move any of the BMDS sensors and weapons systems to new locations at the start of the simulation. This allows for different configurations of defensive capabilities to be tested in the simulation. The missile engagement process is displayed in a three-dimensional space. The missile flights are observable from their launches until destruction. The destruction of the incoming missiles occurs either by reaching their intended targets or by getting destroyed by the BMDS interceptors.

A number of metrics are displayed to the user during runtime and are collected for analysis at the end of the simulation run. The key outputs include which sensors detect, track, classify, and discriminate the specific incoming ballistic missiles during the simulation and the time that each of these functions takes place. The weapon system that is used for an engagement is also captured along with the outcome of that engagement process. The number of missiles that each of the BMDS elements are tracking is also captured over the course of the simulation. This allows for the user to analyze which BMDS elements were the most utilized during the simulation based on the starting location of the enemy missiles. The user can then test new configurations of the scenario to determine how to improve the engagement process and destroy all of the enemy missiles while expending the smallest number of interceptors.

A sample display of the simulation is shown in Figure 4. The United States Central Command (USCENTCOM) is the AOR being used in Figure 4. The incoming ballistic missiles change color to display to the user whether the missile has been detected and not tracked, detected and tracked, or not detected or tracked. The locations of the BMDS elements are clearly visible on the map. Various metrics are displayed around the view area. The number of threats, the number of interceptors launched, and the number of successful interceptions are shown along the top of Figure 4. The bottom of the figure displays the number of ground-based, sea-based, and space-based sensors and the total number of threats that are being tracked during the current time step of the simulation. This provides the user a clear representation of what is occurring in the simulation at any point in time.

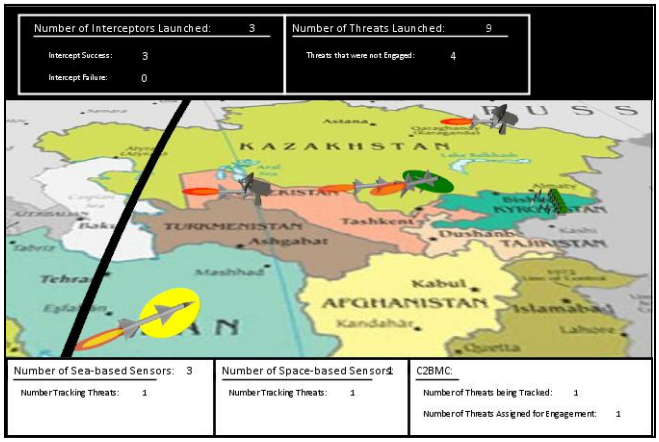


Figure 4: Example Simulation using USCENTCOM AOR

7. CONCLUSION

This model and simulation provide a proof-of-concept that the agent-based modeling paradigm can be used to represent the BMDS. The use of the MS-SDF allows for all of the needed components within the BMDS to be identified and for the characteristics and functions of each element to be identified within a reference model. The reference model also contains the requirements provided for the model. The MS-SDF also allows for a transfer from gathering what is known and assumed about the BMDS to creating a conceptual model that can easily be checked for consistency against the reference model. The simulation is constructed by following the "plan" that the conceptual model provides for answering the modeling questions. The verification and validation process for the simulation is easier to conduct because the reference model, conceptual model, and simulation are all explicitly expressed in a textual and graphical format by following the MS-SDF. This allows the SME, the modeler, and the client to follow the creation of the simulation from the very beginning of the modeling process.

ABM allows for an intelligent software agent to represent the human decision-making component that comprises the BMDS. The behaviors and attributes of the agent that lead to making a decision are embedded within the structure of the agent after being clearly defined in the conceptual model and verified by the SMEs. The use of agents provides an effective method for representing the layered missile defense system that is created by the BMDS. This is possible through the representation of each of the different BMDS elements as agents that are represented with their specific capabilities and functions within a scenario. The simulation can test different sets of ROEs, communication capabilities, and configurations of BMDS elements in order to determine which configurations have the greatest chance of success in defending against incoming ballistic missiles.

REFERENCES

[1] Jay, E. F. "C2BMC for Ballistic Missile Defense." *Military Space & Missile Forum*, 2(5): 8-10, September/October 2009.

[2] Tolk, A., S. Diallo, J. Padilla, and H. Herencia-Zapana. (2013). "Reference modeling in support of M&S – foundations and applications," *Journal of Simulation.* Advance online publication. Doi: 10.1057/jos.2013.3.

[3] Strickland, J. *"Missile Flight Simulation (2nd Edition),"* Lulu Publishing, Simulation Educators, May 2012.

[4] Ender, T., Leurck, R. F., Weaver, B., Miceli, P., Blair, W.D., West, P., and Mavris, D. "System-of-Systems Analysis of Ballistic Missile Defense Architecture Effectiveness through Surrogate Modeling and Simulation," *IEEE Systems Journal* 4(2):156-166, June 2010.

[5] Niland, M.W. "The migration of a collaborative UAV testbed into the FLAMES simulation environment." *Proceedings of the Winter Simulation Conference*, pp. 1266-1272, December 2006.

[6] Bharathy, G.K., Yilmaz, L., and Tolk, A. "Agent Directed Simulation for Combat Modeling and Distributed Simulation," in *Engineering Principles of Combat Modeling and Distributed Simulation* (ed. Tolk, A.), pp. 669-713, John Wiley & Sons, March 2012.

[7] Garrett, R. K., S. Anderson, N. T. Baron, and J. D. Moreland, Jr. "Managing the Interstitials, a System of Systems Framework Suited for the Ballistic Missile Defense System" *INCOSE System Engineering Journal*, Wiley On-Line Publications, October 2010.

[8] Russell, S. J., & Norvig, P. (2003). *Artificial intelligence : a modern approach* (2nd ed.). Upper Saddle River, N.J.: Prentice Hall/Pearson Education.

[9] MDA. (2012). *The Ballistic Missile Defense System*. Missile Defense Agency Fact Sheet.

[10] MDA. (2012). *Cobra Dane Upgrade*. Missile Defense Agency Fact Sheet.

[11] MDA. (2012). *Upgraded Early Warning Radars, AN/FPS-132*. Missile Defense Agency Fact Sheet.

[12] MDA. (2012). *Command, Control, Battle Mangement, and Communications*. Missile Defense Agency Fact Sheet.

[13] Erriksson, H. E. (2004). *UML 2 toolkit*. Indianapolis, IN: Wiley Pub.

BIOGRAPHIES

Christopher J. Lynch is a Graduate Research Assistant at the Virginia Modeling, Analysis and Simulation Center at Old Dominion University. He received a BS in Electrical Engineering in 2011 and a MS in Modeling and Simulation in 2012 from Old Dominion University. His research interests include multi-paradigm modeling and agent-based modeling with a large number of agents.

Saikou Y. Diallo is Research Assistant Professor at the Virginia Modeling, Analysis and Simulation Center at Old Dominion University. He received a B.S. in Computer Engineering, and a M.S and Ph.D. in Modeling and Simulation from Old Dominion University.

Andreas Tolk is Professor of Engineering Management and Systems Engineering with a joint appointment to Modeling, Simulation, and Visualization Engineering at Old Dominion University. He is also affiliated with the Virginia Modeling Analysis and Simulation Center. He holds a M.S. and Ph.D. in Computer Science from the University of the Federal Armed Forces in Munich, Germany.

VALIDATION REPORT OF COMPOSER PROPAGATION SIMULATION: CASE STUDY OF CAMP BLANDING, FL

Nicholas Wright and Robert Kewley (Advisor)
United States Military Academy

INTRODUCTION

In assessing the costs, performance, and risks of fielding a deployable 4G LTE communications capability for the Army's Android-based dismounted command and control system, Nett Warrior, it is important to be able to model how the cellular network will function in multiple operational scenarios. Computer models, in addition to real world simulations, can provide the information necessary for a good decision to be made regarding a variety of factors: the minimum infrastructure needed to provide reliable support throughout a mission; the performance characteristics of the system under particular operational demands; optimal component design, such as antenna type and receiver capabilities; and predicting feasibility in unfavorable conditions. Computer models are preferable to real world simulations because they save money, time, and are flexible to model a variety of scenarios across multiple environments – even ones that a researcher may not have physical access to.

Because geography and the environment are important factors in the performance of wireless communication, if a cellular network is going to be implemented as the backbone to the Nett Warrior system it should be capable across the diverse locations that units will operate. In order to ensure its functionality, and not risk the lives of the soldiers on the ground, decision makers need to be capable of modeling how the system will perform and outfit the unit with the proper infrastructure to ensure operability. This paper attempts to validate and recommend changes to a computer model that has this capability.

A. PURPOSE

Understanding propagation characteristics of a cellular communications network is important for finding the optimal employment of the system. Variables such as the number and disposition of base stations, antenna types, and power requirements are configured to optimally satisfy the link budget, which, depends heavily on correct inputs from propagation path loss data. In a text covering LTE Network Design, Korowajczuk lays out five steps concerning network design: market modeling, network strategy, network design, network optimization, and performance assessment. In the network design phase, a rigorous field measurement campaign precedes the establishment of propagation models and parameters that are used as a basis for the optimization in the fourth step of the design phase (Korowajczuk 516). The significance of this becomes apparent when the military application of cellular network planning is compared to the industry norm. Military networks will not have the luxury of surveying the operational area in the same capacity that an industry provider may scope a city prior to implementation. Therefore, the propagation models used to find the optimal network design in a military setting, whether for reliability or bandwidth, will have to depend on robust models that are accurate in a variety of environments without relying on model and parameter adjustment after field data is collected.

A working network planning model that has been made available to the Army through Lockheed Martin is the cellular modeling software named the Communications Planner for Operational and Simulation Effects with Realism (COMPOSER). COMPOSER has been shown to have faster computation times when compared to a similar defense model referred to as the Joint Communications Simulation System (JCSS); this advantage arises from its different methodology in computing network performance, relying more on statistics of past behavior than pure mathematical models (Drexel 2009).

While verification of the COMPOSER model has been completed internally, validation of the model across the many possible environments it may encounter will continue to be assessed and recalibrated. Camp Blanding, FL is a military facility that many technology venues use to test equipment in an environment that closely resembles South American operations. The plain topography is not ideal for testing terrain based fading, but the monotonous nature allows for the control of a variable of interest concerning the effects of vegetation on signal strength. The hope of this report is that the results of this model validation isolating the foliage effect will result in useful parameter adjustment for future models in similar environments.

B. RF PROPAGATION PRINCIPLES

When considering propagation models, Seybold extends a useful warning to his readers in the introduction of his text:

> RF propagation modeling is still a maturing field as evidenced by the vast number of different models and the continual development of new models. Most propagation models considered in this text, while loosely based on physics, are empirical in nature. Wide variation in environments makes definitive models difficult, if not impossible, to achieve except in the simplest of circumstances, such as free-space propagation.(Seybold 21)

In the context of cellular networks, near earth propagation models have more effect than those intended for satellite transmissions or radar applications. When the original effective isotropically radiated power (EIRP) is generated at the base station from the effective radiated power (ERP) of the transmitter combined with the antenna gain, a significant portion of the original EIRP is lost by the time it reaches the receiver (Equations 1 and 2). This loss is referred to as the path loss of the signal, and is due to multiple types of fading. Masihpour classifies fading into three categories: reflection, diffraction, and scattering. Of these three behaviors, scattering has the worst effects on signal viability because of the narrow scope of the receiver's view. Scattering is most likely to occur when the signal is interrupted by objects that are relatively the same size of the wavelength, thus at a frequency of 700 MHz objects around half a meter in size fall into this category.

$$EIRP(dB) = P_T - L_{connector,\ cable} + G_T$$

Equation 1. (Masihpour 108).

$$PR = PT + GT - LP + GR - AM$$

Where

AM(dB), (L*T* +L*R* in equation (4.2)) represents all the attenuation losses such as feeder loss, link margin, diffraction losses, losses due to mobility (Doppler), and the effects of rain, trees and obstacles in the signal path.

GT (dBi) is the transmitter antenna gain

GR(dBi) is the receiver antenna gain

PR(dBm) is the received power at the receiver

PT (dBm) is the transmitted power

LP (dB) is the path loss in the physical medium between the transmitter and receiver.

Equation 2. (Masihpour, 105-6).

The Longley-Rice and TIREM models that are primarily used in COMPOSER account for all three of these fading effects based off of detailed input data of the terrain. While topography information is readily available from national databases such as the NASA Shuttle Radar, foliage data is much harder to acquire and input into the model. Some rough estimates of vegetation effects are available in Weissberger's model (Equation 3) and the International Telecommunications Union (ITU) recommended model (Equation 4). Seybold points to both of these equations as the better-known foliage models but prefaces them with a familiar tone of caution suggesting that, "it is valuable to verify a particular model's applicability to a given region based on historical use or comparison of the model's predictions to measured results" (Seybold 135).

$L\{dB\} = \begin{cases} 1.33F^{0.268}d_f^{0.588} & 14 < d_f < 400\ m \\ 0.45F^{0.284}d_f & 0 < d_f < 14\ m \end{cases}$ Where d_f is the depth of foliage along the LOS path in meters[1] F is the frequency in GHz	$A_{ev} = A_m\left[1 - e^{\frac{d_y}{A_m}}\right]\ dB$ Where d is the length of the path that is within the woodland in *meters*[2] *y* is the specific attenuation for very short vegetative paths (dB/m)[3] A_m is the maximum attenuation for one terminal within a specific type and depth of vegetation (dB)[4]
Equation 3. Weissberger Model	**Equation 4.** ITU Model

II. METHODOLOGY

A. TESTING PLAN

This research team developed an adjusted testing plan to fit the parameters of the network available at the time of testing. Originally, the scope of the testing environment encompassed multiple kilometers of terrain with a high powered transmitter. Due to the constraints on the system, the testing environment was limited to a 700 meter radius which was not ideal for comparing data to the useful range of the COMPOSER propagation algorithms meant for a distance of at least 1 kilometer (Seybold 143).

The data collection process consisted of two operators and one data recorder. Either operator would report a signal strength reading in decibels and the data recorder would record the signal strength along with the latitude and longitude readings from the device GPS which were recorded to five decimals of accuracy. Readings were taken from an Android device in the North, South, and West directions including line of sight locations to the base station as well as heavy vegetation. The maximum recorded distance extended to 765 meters away from the base station antenna. Sixteen data points were taken in vegetation and eighteen points were taken out of vegetation. All points were recorded under similar temperature and atmospheric conditions as they were taken on the same day within a five hour time interval. A graphical layout of the testing points can be seen in the illustration below (Figure 1).

[1,2] Both distances of vegetation obstruction (in meters) were estimated by dragging the Google Earth ruler tool from the start of the vegetation in the path to either the data collection point or the end of the last vegetative patch between the tower and the point.

[3] Values for specific attenuation can be found in the graph included in Appendix A

[4] The ITU model could not be generated because the maximum attenuation values were not available

Figure 1. Testing Map. 7 February 2013.

After the collection process was completed, each GPS location was inputted into the COMPOSER model and run under the TIREM propagation model updated with the Camp Blanding topography and foliage information. The projected signal strength according to the model was then compared to the actual signal strength recorded in the experiment.

B. DATA ANALYSIS

At a glance, COMPOSER appears to track the general trend of the actual attenuation when the model is adjusted for its overall mean error of 40 decibels. That is, every predicted value from the COMPOSER model is lowered by 40 decibels (Figure 2). Unfortunately, while the predicted values track the general trend, they do not closely tack the individual deviations from the general distance fading caused by vegetation or lack of vegetation (points 19, 20, 21, 34, 35).

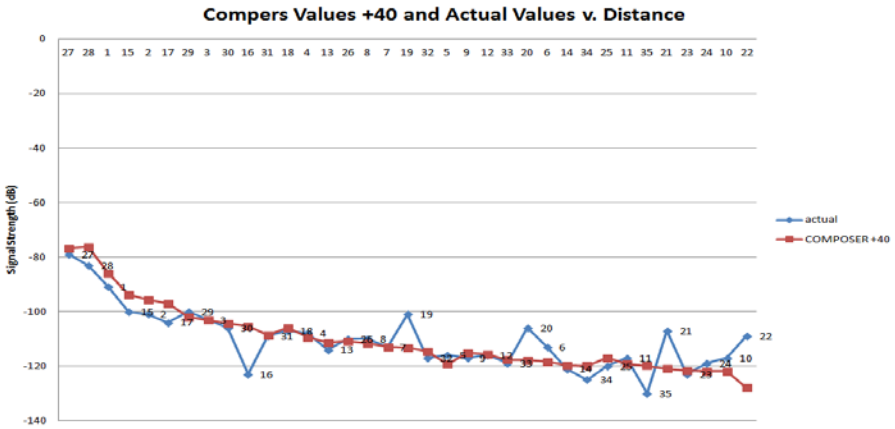


Figure 2. The COMPOSER estimates were found to be generally close to the actual observations when 40 dB were subtracted from the estimated signal strength.

When the model was stripped of its foliage modeling and recomputed for each data point on the map, a paired Student's t-test gave a 65% likelihood that both runs came from the same underlying population. This is not enough uncertainty to reject the notion that both distributions came from the same population. Because each point is a function of multiple factors (mainly distance and vegetation), a correlation coefficient is more representative model's accuracy. Table 1 shows a comparison between the COMPOSER results with and without the foliage model using both the paired t-test and Pearson's correlation coefficient. Table 2 shows identical metrics for the points that are in question. Comparing Table 1 to Table 2 suggests that COMPOSER is good at tracking the general trend of all the data points, but not good at tracking the extreme cases of points 19, 20, 21, 34, and 35 which either incurred large distances of vegetation or large distances with line of sight.

COMPOSER compared to Actual Data	**Foliage Model?**	
	No	**Yes**
Paired t-test	0.919976	0.9994
Correlation Coefficient	0.838931	0.837982

Table 1. Results for entire data set

COMPOSER compared to Actual Data	**Foliage Model**	
	No	**Yes**
Paired t-test	0.40774	0.414524
Correlation Coefficient	0.597003	0.596994

Table 2. Results for specific points in question 19, 20, 21, 34, 35 (clearly vegetative and clearly line of sight)

Comparing the two tables above raises the question of whether there exists a better foliage model for the environment of Camp Blanding, FL. When applying the traditional Weissberger model path losses to the COMPOSER data points that do not include a foliage adjustment, the correlation for all data points dropped to 0.76 while the correlation for specific data points increased to 0.99. Table 3 shows how adjusting one exponent parameter of the Weissberger model to achieve optimal correlation resulted in an equal coefficient for all points (0.835) and a superior coefficient for the specific points (0.984).

COMPOSER with Weissberger Adjustment	**Data Points**	
	All	**Specific (19,20,21,34,35)**
Normal Equation	0.76	0.99
Adjusted Exponent	0.835	0.984

Table 3. Results for incorporating the Weissberger model for vegetation path losses to the COMPOSER model that does not all ready adjust for vegetation

III. CONCLUSION

The outputs from COMPOSER correlated well with the data set as a whole, visually it is evident that the general trend due to distance fading was matched by the COMPOSER model (Table 1). However, the simulation did not correlate well with some of the points in the data set that were exposed to either extreme of vegetation or line of sight (Table 2). When the Weissberger model was substituted for COMPOSER's organic model, significant gains were realized in the correlation with the extreme points but it sacrificed correlation with the data set as a whole.

To arrive at a recommendation, the Weissberger parameters were optimized to match the field data from the Camp Blanding environment. If the long distance vegetation exponent parameter of 0.588 in Equation 3 is modified to a value of 0.404[5], the correlation coefficients increase to 0.835 for all data points and 0.984 for the extreme points. With this modification to the foliage model, COMPOSER could track the changes in vegetation significantly more effectively without sacrificing its tracking of the general trend.

This finding however, highlights the problem of hindsight in modeling foliage path loss effects that Seybold mentions. For military applications, this is particularly frustrating because the opportunity to adjust the model based off of field data before the network is deployed is non-existent. In order to properly mitigate this risk, it will be important for the military communications analyst to have a large data base of pre-calculated coefficients that will provide him with a satisfactorily accurate model of the environment the network will be deploying in.

REFERENCES

Allen, Frank, Greg Tarancon, and Nick Wright. Signal Strength Readings. 8 Feb. 2013. Raw data. Range 1, Camp Blanding.

Korowajczuk, Leonhard. *LTE, WIMAX, and WLAN Network Design, Optimization and Performance Analysis*. Chichester, West Sussex, U.K.: Wiley, 2011. Print.

Modeling and Simulation in Support of COMPOSER and CJSMPT: COMPOSER and JCSS Comparison. Rep. Vol. Phase 8. Camden: Drexel University, 2009. Print. Applied Communications and Information Networking.

Masihpour, Mehrnoush. "WiMax and LTE Link Budget." *Planning of WIMAX and LTE Networks*. Ed. Johnson I. Agbinya. N.p.: n.p., n.d. 105-35. Print

Seybold, John S. *Introduction to RF Propagation*. Hoboken, NJ: Wiley, 2005. PDF.

[5] This value was arrived at by the Excel solver.

APPENDIX A

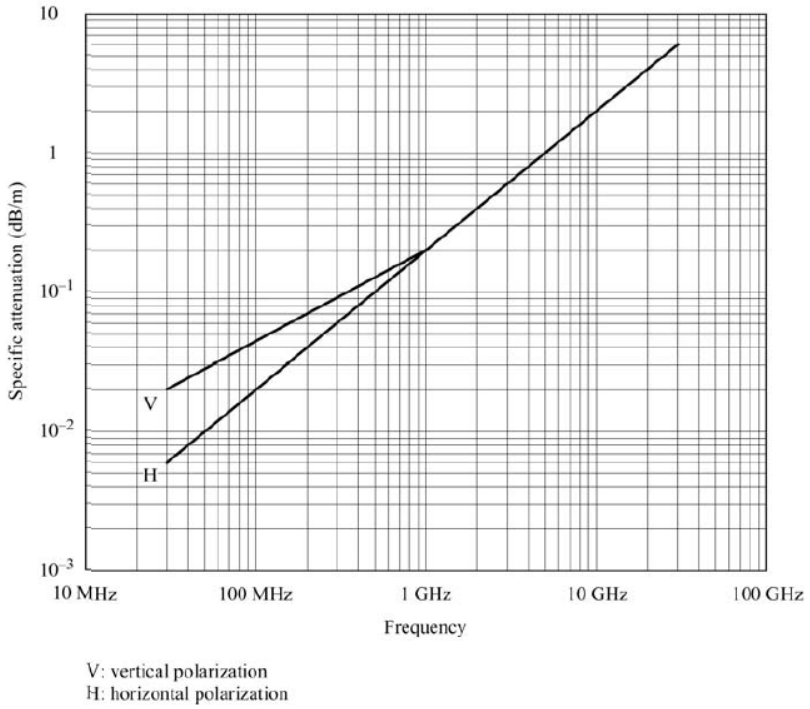


Figure 3. Specific Attenuation by frequency for ITU foliage mode coefficient *y.*

Shiwei Jiang

International Studies, ODU

Sjian002@odu.edu

Game Theory and the North Korea Nuclear Puzzle

Abstract: North Korea is currently a hot topic in international affairs. This paper probes into the North Korean nuclear issue through game theory. It demonstrates that the four-element analysis of game theory (players, strategies, rules and information) is an effective approach to analyze the North Korean nuclear puzzle. To better design a mechanism solving the North Korean nuclear issue, scholars can use the four-element analytical approach in future.

Key words: North Korean nuclear issue, international affairs, game theory, four-element analysis, security mechanism

Organization

1. Main task: The major task of my research is to analyze the North Korean Nuclear issue after the Cold War through game theory perspective.

2. Purpose: Why is it feasible to analyze North Korean Nuclear Issue with game theory?

Because I found that various concepts in game theory can be applied to the North Korean nuclear issue, such as credible threats, brinkmanship, to dominant strategies, Tit for Tat strategy and so on. [1]

A. Credible threat: How credible are the threats of sanctions from the US/UN, and North Korea of waging war?

B. Brinkmanship: when a player pushes a dangerous situation to the brink of disaster, in the hope that the other player will compromise. (e.g., several nuclear tests in recent years, launching so many test missiles, sinking South Korean ship and bombing South Korean Island in recent months).

C. Dominant strategy: given that the US already has nuclear weapons, North Korea's rational response (indeed, its dominant strategy) is to also build nuclear weapons. The problem (as in the classic Prisoners' Dilemma) is the incentive to "cheat" or "defect" is too much. As Kim Jong Il said, "Our military first policy calls for an eye for an eye, a tooth for a tooth, retaliation for retaliation, ultra-hardline for hardline, war for war, total war for total war, nuclear war for nuclear war".

[1] http://tutor2u.net/blog/index.php/economics/comments/game-theory-and-north-korea/

D. Tit for Tat strategy: Kim Jong-Il's quote "*Our military first policy calls for an eye for an eye, a tooth for a tooth, retaliation for retaliation, ultra-hardline for hardline, war for war, total war for total war, nuclear war for nuclear war*." [2]

3. Literature: Numerous studies have probed into the North Korea nuclear issues and game theory separately. However, few of them carefully combine both. Games become complex, as there are multiple players intertwined with complicated interests. Francis Fukuyama's "Re-Envisioning Asia" depicts that in East Asia there are many pairs of ties: U.S.-Japanese Security Treaty, ASEAN, APEC, SCO, and U.S.-South Korean Alliance. Besides, "Asia lacks strong multilateral political institutions. Europe has the EU and NATO", which creates vacuum for conflicts and makes the North Korean nuclear game more complex. Journal of Conflict Resolution, International Journal of Game Theory, International Security and International Study Quarterly are prestigious journals that cover most articles about North Korean nuclear issue and game theory. Some prominent works concerning global security and game theory are Thomas Schelling's "The Strategy of Conflict" and "Arms and Influence", Robert D. Putnam "Diplomacy and Domestics: The Logic of Two-level Games", Kenneth Waltz: Theory of International Politics, Robert Jervis: "Cooperation under the Security Dilemma" and "War and Misperception", Bruce Bueno de Mesquita: "The Predictioneer's Game", and James E. Dougherty and Robert L. Pfaltzgraff's "Contending Theories of International Relations."

4. Hypothesis:

A. To explain North Korean Nuclear issue, a better way is to combine international relation theory, deterrence theory and analytic tool.

[2] http://www.atimes.com/atimes/Korea/KF12Dg01.html

B. North Korean Nuclear issue is a multiple level complex game. The more actors get involved, the more complexity it has.

C. Negotiation players, their strategies, rules and information matter. The change of these elements of game will change the status quo

5. The Outline of major sections:

Section 1: What leads to the complexity of North Korea Nuclear issue?

----cultural explanation, mainstream IR theory explanation, deterrence theory explanation, Sagan's three theoretical models and finally game theory explanation, the nuances between different explanations.

Section 2: What are the types of games?

"The Nuclear Arms Race: Prisoner's Dilemma"[3] ----- a game of defection. No possibility to cooperate. Defection is more attractive than cooperation.

Stag-hunt Game ---- assurance game**.** There is a trust dilemma in Stag-hunt game. There is possibility to cooperate. Choosing cooperation, you may get nothing or get a lot of benefits. Not to cooperate, you can get something at least. Both sides need assurance.

Chicken Game ---- the game of coward. Who evade first when two cars rush into each?

Section 3: How does the game work? The game works based on four-element framework: the players, strategies, rules and information in North Korean nuclear issue.

6. Conclusions: (implication and future research)

[3] S. Plous: The Nuclear Arms Race: Prisoner's Dilemma or Perceptual Dilemma? *Journal of Peace Research*, Vol. 30, No. 2. (May, 1993), pp. 163-179.

A. What implications can we draw from the analysis by using game theory? (e.g. what are the possible paths that may lead to a peaceful resolution?)

B. What is the limitation of using game theory to analyze the North Korea nuclear issue, or, what further research needs to be done?

Main Sections:

Section 1: What leads to the complexity of North Korea Nuclear issue?

To explain the complexity of North Korea nuclear issue, there are a set of explanations. First, a basic explanation is historical and cultural theory.

Hunting's cultural clash theory----The conflicts between states are driven by cultural and historical differences. The future clash will be between three civilizations, namely, the East, the West and the Middle East as they represent three distinct cultures.

1. **IR mainstream theories**

a. **Alexander Wendt's social constructivism**----states construct their enemies or friends through their own value, norms and perception. A familiar example is why the U.S. worries more about the several bombs of North Korea compared hundreds of nuclear bombs owned by Britain, not because of national interest, but it is more because of their constructive perception of enemy and friend. North Korea nuclear issue is complex, because mutual constructive perception is complex between the involved players.

b. **Kenneth Waltz's realism perspective**----National interests and security are the priority in North Korea nuclear issue. The core of North Korea nuclear issue is states trying to gain national interests and protect their security.

c. **Keohane and Nye's liberalism perspective** does not directly address the North Korean nuclear issue. However, they meticulously discuss interdependence. Actually, it is the interdependence between the six players, particularly China and America, that maintains the security stability in East Asia so far.

2. Deterrence theory----explains the motivation of North Korea's pursuit of nuclear weapons, U.S. extended deterrence in East Asia, and more importantly can better explain the security stability in East Asia as three major nuclear powers (China, the U.S. and Russia) are clustered in the region.

3. **Sagan's three theoretical models**: the security model, the domestic politics model, and the norms model. This theory actually has some overlaps with the above explanations. The security model is overlap

4. **Finally, Game theory**---better explain the process and strategies conducted by six players in the North Korean Nuclear issue.

Section 2: what type of game fits the North Korean nuclear issue?

The pursuit of relative power is often described as zero-sum game, where one state's gain in power comes at another state's loss. Game practitioners and game theorists have recognized that the characteristics of games with more than two players are different from those of two-player games. Chief among these is the presence of the coalition dynamic.

The U.S. rejects bilateral talk or direct talk with North Korea. For both America and North Korea, they have two options, choosing direct talk or no direct talk. In reality, The U.S. rejects bilateral talk or direct talk with North Korea. So, we can expect that for the U.S., the payoff of no direct talk with the North Korea is higher than the payoff of direct talk with North Korea. By contrast, North Korea is very actively pushing America to have a direct bilateral talk directly. Thus, we expect that for North Korea, the payoff of direct talk with America is higher than the payoff of no direct talk with America. There are numerous games in game theory that have been used to analyze nuclear issues. Multi-players makes the North Korea Nuclear issue complex (6 players at least). Which type of games does North Korea nuclear issue fit?

A: **The Nuclear Arms Race: Prisoner's Dilemma**

This game is usually used in the case of two states with paralleled strengths; and the best strategy for both sides is to defect instead of cooperating. Thus, prisoner's dilemma is repeatedly used to analyze the Soviet-U.S. deterrence in the Cold War. Both sides seldom considered cooperation. It is also used to analyze the economic conflict between China, Japan and America in recent years. For North Korea, this game may not be that suitable, because the U.S. and North Korea do not math in their strengths. Besides, there are other players in this game. This game is a more two-level game rather than a multi-level game. [4]

Chicken Game

In previous literature, chicken Game was adopted to analyze Cuba Crisis in many scholarly publications. Gilbert Funabashi Yoichi interprets 'the Korean peninsula's nuclear crisis as

[4] See, S. Plous's "The Nuclear Arms Race: Prisoner's Dilemma or Perceptual Dilemma?", Schelling's "The Strategy of Conflict" and "Arms and Influence", Robert Jervis's "Cooperation under the Security Dilemma," and Bruce Bueno de Mesquita's "The Predictioneer's Game".

continuing the U.S.-North Korean game of chicken. Gilbert Rozman's "The North Korean Nuclear Crisis and U.S. Strategy in Northeast Asia" puts that "The U.S.-North Korean game of chicken to see which side would blink first continued."

Chicken game is also related to brinkmanship. The examples are a series of North Korea's provoking actions from the 1990s up to date. In the Chicken game, North Korea is bold and attempts to prove that he is not the coward. Instead, as the U.S., South Korea and Japan have restrained their behavior over years, they seem to be coerced and passive in this nuclear game.

Stag-hunt Game

In this game, what parties need is nothing but assurance. For instance, the Six Party Talks attempts to create a platform where the involved states could give the signal of assurance to each other. Besides, the North Korean nuclear issue is chiefly based on the economic and nuclear assurance between China and the U.S. If the assurance is reliable and attractive, the parties are willing to cooperate and finally can gain tremendous benefits, "a big stag".

Section 3: How do powers game with each other on the North Korea nuclear issue?

1. players---"six players: the U.S., China, Russia, Japan, South Korea and North Korea"

Source: Yinhay Ahn "A Survival Game: China, Japan and North Korea in 2002. Francis Fukuyama "Re-Envisioning Asia" Five-power forum---"Five-Power forum would be particularly useful in dealing with several foreseeable problems." Jessie Bernard: "Parties and Issues in Conflict". Joseph Cirincione's "The Asian Nuclear Reaction Chain". The North Korean nuclear issue game is intertwined with a set of bilateral games, such as the military game between China and America, the territory game between China, Russia, Japan, South Korea and so forth. Other

actors beyond the 6 players are Iran and Pakistan. Iran and Pakistan's linkage with North Korea also strengthens the complexity of North Korean nuclear issue. Although there are a bunch of players in the game, we should understand that the core of North Korean nuclear issue is security-oriented and that the key players in the game are China and the U.S.

2. Their strategies (cooperate or defect) and **corresponding payoffs** (comparing gains and losses of different strategies),

The U.S. strategy: (see table)

North Korea strategy: (see table)

Table: [5]

Predicting the Outcome of the North Korean Nuclear Crisis
by Brian Gongol

Game theory provides a useful method for predicting the likely outcomes of future events. With North Korea threatening to test a long-range nuclear weapon, it is useful to apply this method in order to predict the possible outcomes. The first action is shown in the left-hand column, so each of the following starts with North Korea's first move, followed by the possible US responses:

	US launches successful counter-measures	US launches unsuccessful counter-measures	US does nothing	US launches preemptive airstrikes with UN approval	US launches preemptive airstrikes without UN approval
NK launches successful test	US establishes that it has an effective deterrent US unlikely to engage in combative	US reputation damaged Gives NK pretext to launch aggressions against SK	NK attracts diplomatic attention and gravity equal to that being paid to Iran	(not applicable)	(not applicable)

[5] http://www.gongol.com/research/predictions/2006/northkorea/

	retaliation US likely to approach UN for approval to launch airstrikes				
NK launches successful attack against foreign target	Combat ensues with massive retaliation US forces already spread thin due to engagements in Iraq, Afghanistan	US reputation damaged Combat ensues with massive retaliation US forces already spread thin due to engagements in Iraq, Afghanistan	Virtually impossible to conceive	(not applicable)	(not applicable)
NK launches unsuccessful test	US establishes that it has an effective deterrent NK may try to seize pretext for aggression against SK US likely to approach UN for airstrikes against NK nuclear targets Kim Jong-Il may conduct a purge of his domestic and military leadership	US reputation damaged US less likely to pursue airstrikes against NK nuclear targets than if test had been successful Kim Jong-Il may conduct a purge of his domestic and military leadership	NK forces other nations to negotiate and/or supply aid US may approach UN for approval to launch airstrikes Kim Jong-Il may conduct a purge of his domestic and military leadership	(not applicable)	(not applicable)
NK does nothing	(not applicable)	(not applicable)	NK forces other nations to negotiate	Better chance of success than if delayed	High probability that NK will use human shields,

	and/or supply aid	by wait for UN approval More difficult for US to win allies than if UN approval is granted May damage US reputation May sufficiently destabilize Kim Jong-Il to lead to internal power struggles and possible coup	since transparency of a UN debate would give NK time to organize May sufficiently destabilize Kim Jong-Il to lead to internal power struggles and possible coup

Other four players' strategies:

China: make use of North Korea nuclear issue to negotiate with the U.S.. At the same time, China curbs North Korea's nuclear development by economic pressure.

Japan: Strongly against North Korean nuclear development. Wish the U.S. could harsh on North Korea. Suspend economic relation with North Korea

South Korea: Different from Japan, South Korea keeps certain economic relation with North Korea, still has the sympathy and the feeling of compatriot toward North Korea. Need the U.S. extended nuclear deterrence.

Russia: make use of North Korea nuclear issue to negotiate with the U.S.. involve nuclear technology in North Korea issue.

3. Rules---Designing a mechanism for Multi-lateral security cooperation

a. Static, dynamic; single game or iterated,

b. Different rules: A. for example, currently, Obama's strategy is the U.S., China, Japan and Russia talk first and then the five negotiate with North Korea. B. 6 players talk with each other at the same time. C. The U.S., Japan, South Korea talk first, Russia, China and North Korea talk first. And then the two groups talk with each other. D. North Korea and the U.S. talk first, and then they talk with other players). Gilbert Rozman puts that "once Six-Party Talks were proceeding; their goals could have been approached with urgency and realism on how to establish a consensus of 5 vs. 1 over a broad agenda." [6]

c. Designing a mechanism for Multi-lateral security cooperation[7]

4. Information

a. lack of information easily leads to misperception. Misperception leads to conflicts. Conflicts lead to war. [8]

b. states are more willing to cooperate in a game, when they know more about each other. Robert Jervis: War and Misperception[9]

c. a better mastery of your opposite's information is a source of your power in war. [10]

[6] Gilbert Rozman: "The North Korean Nuclear Crisis and U.S. Strategy in Northeast Asia", *Asian Survey*, Vol. 47, No.4, 2007.
[7] See, Peter Van Ness: "Designing a Mechanism for Multi-lateral Security Cooperation In Northeast Asia", Young Whan Kihl: "Unraveling of U.S.-DPRK Nuclear Accord? A Post-Mortem Analysis of the Six-party Talks(SPT) Process" Jessie Bernard: "Parties and Issues in Conflict", *Journal of Conflict Resolution*, Vol. 1, No. 2, 1957.
[8] Robert Jervis: "War and Misperception", *Journal of Interdisciplinary History*, Vol. 18, No. 4, 1988.
[9] Jessie Bernard: "Parties and Issues in Conflict", *Journal of Conflict Resolution*, Vol. 1, No. 2, 1957.
[10] William Reed: "Information, Power and War", *American Political Science Review*, Vol. 97, No. 4, 2003.

Conclusion

Section 1: Implication

a. The North Korea is a complex multi-players game. Therefore, to analyze the North Korean nuclear issue, we should unfold around the four elements of game (players, strategies, rules and information).

b. the three game models (prisoner dilemma, stag-hunt and chicken game) tell us that states are rational actors and they pursue their benefit and interest. To defect or cooperate depend on payoffs, the reliability of assurance, and information.

Source: Jaewoo choo: "Strategic implications of Six-Party talks for East Asia Future Security", James H. Lebovic: "The Law of Small Numbers: Deterrence and National Missile Defense"

c. The current Six-Party talk is not enough. New mechanisms are needed. To design a mechanism solving the North Korean nuclear issue in future, scholars can follow the four-element analytical path.

Section 2: Limitation

Rationalism (rational choice theory) and deterrence theory

a. The limitation of rationalism: Are states always rational or benefits maximizer? If states policies are driving by irrational leaders, game theory and deterrence theory do not hold the water.

b. the limitation of deterrence theory: Deterrence works between China, Russia and America. But it is questionable whether deterrence is effective between North Korea and South Korea or between China and Taiwan. Some hold that since North Koreans and South Koreans, or Chinese and Taiwanese are the same people and they were separated not long ago. So, the public and even some decision-makers are against using nuclear weapons toward their own people. For this matter, when war happens, it is possible for North Korea to unclearly attack Japan. But it is not likely that North Korea attack the South Korea by nuclear weapons. The same to China, if there is a nuclear war, it is more likely to nuclear attack Japan and the U.S. rather than Taiwan at first, because the general public and Chinese leaders may not accept nuclear bombing their own people. This thought is more related to sociopsychological approach in political science.

Section 3: Areas of Future Research

1. The effects of Iraq and Libya on North Korea nuclear issue

The recent Libya war and the Iraq war in 2003 seem to have certain effects on North Korean's unwillingness to give up nuclear weapons. What are the effects and to what degree?

2. New security mechanism for East Asia

The East Asia lacks multi-lateral cooperative security mechanisms. The Six-Party talk does not live up to a regional security mechanism as it is not consistent and suspended several times. New security mechanisms for East Asia are needed. What kind of security mechanism can be built in this region and how to distribute the power and rights in new security mechanisms?

3. Is nuclear threat an obsolete idea in this region?

In East Asia, economic conflicts seem to overshadow political and military conflicts. With the deepening economic cooperation between China, the U.S., Japan, South Korea and Taiwan, more and more people in this region talk about economic conflicts, while fewer talk about nuclear threat. To lots of people in this region, the nuclear war in the Taiwan Strait or in the Korea peninsula is quite unlikely in predictable future. Therefore, a new question may arise, is nuclear threat an obsolete idea in East Asia?

Bibliography

Publications:

Francis Fukuyama: "Re-Envisioning Asia", *Foreign Affairs*, Vol84, No.1, 2005.

Peter Howard: "Why Not Invade North Korea? Threats, Language Games, and U.S. Foreign Policy", *International Studies Quarterly,* Vol. 48, No. 4, 2004

Kurt M. Campbell: *The Nuclear Tipping Point: Why States Reconsider Their Nuclear Choices*, Brookings Institution Press, 2004.

Anne Harrington de Santana: "Nuclear Weapons as the Currency of Power: Deconstructing the Fetishism of Force", *The Nonproliferation Review,* Vol. 163, No.9, 2009.

Thomas Schelling: *The Strategy of Conflict,* and *Arms and Influence*.

Robert D. Putnam "Diplomacy and Domestics: The Logic of Two-level Games", *International Organization*, Vol. 42, No. 3, 1988.

Kenneth Waltz: *Theory of International Politics,* McGraw-Hill, 1979.

Samuel Huntington: *The Clash of Civilizations,* Simon & Schuster, 1996.

Alexander Wendt: *Social Theory of International Politics,* Cambridge University Press, 1999.

Robert Jervis: "Cooperation under the Security Dilemma", *World Politics*, Vol. 30, No. 2, 1978.

Robert Jervis: "War and Misperception", *Journal of Interdisciplinary History*, Vol. 18, No. 4, 1988

Bruce Bueno de Mesquita: *The Predictioneer's Game*, Random House, 2009.

Jessie Bernard: "Parties and Issues in Conflict", *Journal of Conflict Resolution*, Vol. 1, No. 2, 1957.

S. Plous: "The Nuclear Arms Race: Prisoner's Dilemma or Perceptual Dilemma?", *Journal of Peace Research*, Vol. 30, No. 2, 1993.

Jaewoo choo: "Strategic Implications of Six-Party Talks for East Asia Future Security", *Tamkang Journal of International Affairs.*

Gilbert Rozman: "The North Korean Nuclear Crisis and U.S. Strategy in Northeast Asia", *Asian Survey*, Vol. 47, No.4, 2007.

William Reed: "Information, Power and War", *American Political Science Review*, Vol. 97, No. 4, 2003.

Joseph Cirincione: "The Asian Nuclear Reaction Chain", *Foreign Policy*, Spring 2000

Yinhay Ahn: "North Korea in 2002: A Survival Game", *Asian Survey*, Vol. 43, No. 1, 2002.

Martin Shbik: "Some reflections on the Design of Game Theoretic Models for the Study of Negotiation and Threats."

James H. Lebovic: "The Law of Small Numbers: Deterrence and National Missile Defense", *Journal of Conflicts Resolution*, Vol. 46, No. 4, 2002.

Websites:

http://www.un.org/News/Press/docs/2006/sc8853.doc.htm

http://www.un.org/apps/news/infocusRel.asp?infocusID=69&Body

http://www.fas.org/nuke/guide/dprk/nuke/index.html

http://www.iaea.org/newscenter/focus/iaeadprk/

http://nuclearweaponarchive.org/

http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB87/

http://www.iiss.org/publications/strategic-dossiers/north-korean-dossier/north-koreas-weapons-programmes-a-net-asses/north-koreas-nuclear-weapons-programme/

http://bx.businessweek.com/north-korean-economy/game-theory-and-north-koreas-nuclear-strategy/14386014134297498501-01658a5a72881753ffa65c2dc2507cd8/

Medical Modeling and Simulation

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

Using a Spatial Task to measure Laparoscopic Mental Workload: Lessons Learned and Initial Results
Author(s): Erik G. Prytz

Dynamic Architecture for Simulation-Based Obstetrics Training
Author(s): Tyrell Gardner, Spencer Lane, Lucas Lezzi, and Tien Nhan

Occurrences of Delayed After-Depolarizations and Triggered Activity in RNC Model
Author(s): Sharmin Sultana, Tanweer Rashid, and Christian Zemlin

Modeling and Simulation of the Shape Recovery of Red Blood Cells
Author(s): John Gounley, and Yan Peng

A Parallel Marching Cubes Algorithm for Extracting Isosurface from Medical Images
Author(s): Jing Xu, and Andrey N. Chernikov

Estimating Lower Bounds on the Length of Protein Polymer Chain Segments using Robot Motion Planning
Author(s):Andrew McKnight, Jing He, Nikos Chrisochoides and Andrey Chernikov

Simulation of the Localized 3-dimensional Reconstruction for Electron Cryo-tomography
Author(s): Dong Si, Hani Elsayed-Ali, Wei Cao, Olga Pakhomova, Howard White, and Jing He

Estimating Lower Bounds on the Length of Protein Polymer Chain Segments using Robot Motion Planning
Author(s):Lin Chen, Kamal Al Nasr, and Jing He

Using A Spatial Task to Measure Laparoscopic Mental Workload: Lessons Learned and Initial Results

Erik G. Prytz

***Abstract* — This paper presents some initial results from a study on measuring laparoscopic mental workload using a secondary task. The secondary task draws on the same visual-spatial mental resources as the laparoscopic task. An experiment was conducted to test the utility of the secondary task. It was shown that the secondary task can successfully distinguish among tasks that impose different levels of workload on novice performers.**

I. Introduction

Performing laparoscopic surgery is more mentally demanding than traditional surgery in part because surgeons must operate in three-dimensional space while viewing a two-dimensional display. Consequently, laparoscopy places significant demands on visual attention and requires a great deal of practice to achieve proficiency.

Presently, there is no standard method to measure the mental demands imposed by laparoscopic surgery. However, Stefanidis and his colleagues have used the secondary task technique to assess mental workload [1, 2]. According to Wickens' multiple resource theory, pools of attentional resources are distinguished by three dimensions: processing stages (perceptual/cognitive and response), processing codes (verbal and spatial), and processing modalities (auditory and vision) with the visual processing modality separated into focal and peripheral channels [3]. Two tasks that draw upon the same pool of resources can interfere with one another and increase mental workload. Thus, a secondary task that competes for the same resources as a primary task should be sensitive to differences in mental workload.

A secondary spatial task was developed that requires the same visual processing needed for judging the position of objects on a laparoscopic display. Thus, our objective was to determine if the new secondary task could distinguish differences in mental workload associated with several different tasks in an FLS laparoscopic box trainer..

Manuscript received February 20, 2013. This work was funded in part by a grant from the Agency for Healthcare Quality and Research (1R18HS020386-01) and support of the Modeling and Simulation Graduate Research Fellowship Program at Old Dominion University. This paper was previously published at the 2013 International Meeting on Simulation in Healthcare by Prytz, Montano, Kennedy, Scerbo, Britt, Davis, & Stefanidis. It received the 2nd place award for Best Research Abstract.

Erik G. Prytz is a PhD student in Human Factors Psychology at the Department of Psychology at Old Dominion University, Norfolk, VA, 23529 USA (e-mail: epryt001@odu.edu).

II. Method

Sixteen undergraduate students with no prior laparoscopic experience were recruited to participate in this IRB approved study. They were asked to perform three primary tasks on a box trainer: 1) tracing the outlines of images on a computerized drawing tablet using a stylus attached to a laparoscopic instrument, 2) the FLS peg transfer task, and 3) the FLS cutting task.

The secondary task presented observers with images of four balls in a simulated tunnel, superimposed at 50% transparency over the primary task display to ensure that both tasks were viewed with focal vision. The images were presented for 300 ms every 4 sec. On half of the presentations, one ball changed its position and participants had to verbally identify those shifts in position. Participants performed the secondary task by itself and in conjunction with the primary tasks. The experiment used a within-subjects design with four task conditions: tracing task, peg transfer task, cutting task, and a baseline measure of the secondary task by itself.

III. Results

The percent correct responses on the secondary task were analyzed with a repeated measures ANOVA and showed a significant main effect, $F(3, 45) = 41.285$, $p < .001$, partial $\eta 2 = .747$. Bonferroni-corrected post hoc tests showed that percent correct scores on the secondary task were lower when performed in conjunction with the tracing task (M =80%, SE = 2.5%), $p < .001$, the peg transfer task (M = 72.4%, SE = 3.3%), $p < .001$, and the cutting task (M = 62%, SE = 3.3%), $p < .001$,compared to performing the secondary task alone (M = 92.7%, SE = 2.0%). Further, secondary task scores were significantly lower on cutting than the peg transfer task, $p = .007$.

IV. Discussion

The results show that the spatial secondary task was indeed sensitive to the mental workload associated with the laparoscopic tasks. Scores on the secondary task declined when performed in conjunction with each laparoscopic task. Further, secondary task performance with the cutting task was poorer than with the peg task, suggesting that the cutting task is much more mentally demanding. These initial results show that the secondary task can provide an objective index of mental workload that can complement traditional metrics of speed and accuracy on laparoscopic tasks. Future research will use this task to identify laparoscopic conditions that deprive surgeons of the spare attentional resources needed to multitask effectively in the OR.

Disclosure

This paper was previously published at the 2013 International Meeting on Simulation in Healthcare by Prytz, Montano, Kennedy, Scerbo, Britt, Davis, & Stefanidis. It received the 2nd place award for Best Research Abstract.

References

1. Stefanidis, D., Scerbo, M.W., Korndorffer, J. R. Jr., & Scott, D.J. (2007). Redefining simulator proficiency using automaticity theory. The American Journal of Surgery, 193, 502-506
2. Stefanidis, D., Scerbo, M.W., Smith, W., Acker, C.E., & Montero, P.N. (2012). Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training: A randomized controlled trial. Annals of Surgery, 255, 30-37.
3. Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science, 3*, 159-177.

Dynamic Architecture for Simulation-Based Obstetrics Training

Tyrell L. Gardner, Spencer D. Lane, Lucas C. Lezzi, and Tien C. Nhan

***Abstract* — Obstetrics training has advanced a great deal, from text-based education, observation, hands-on experience, and now with phantom-based simulation. This project proposes an addition to the current phantom-based simulation training, by providing an immersive, adaptable environment in which the training may take place. Not only does the environment provide the visual sensation of a delivery room, but the audio sensations, interactivity of virtual medical devices, and allowing for more extensive review of student performance. This report delves into these expanded features of phantom-based training simulations from a software development perspective as well as touches on the impact that these enhanced features should provide to medical students training in not only obstetrics, but number of other medical specialties.**

***Index Terms*—Medical Simulation, Virtual Environment, Simulation-Based Training, Obstetrics Training**

I. INTRODUCTION

THE Modeling, Simulation, and Visualization Engineering (MSVE) Capstone Design Team is designing and implementing a realistic medical training simulation for obstetrics residents at Eastern Virginia Medical School (EVMS). The training simulations currently in use at EVMS already use a manikin; however, the surrounding environment does not provide a realistic setting for the student to learn in. In order to enhance the student's learning experience, a realistic training environment is built to surround the manikin. There are interactive virtual devices such as an IV pump, which the student interacts with during the training session. These added components create a more realistic training environment and give the student other avenues of interacting with the training simulation. This project is titled Virtual Labor and Delivery (VLAD).

This paper is divided into nine sections including this introduction. Section Two gives some background information on simulated training, and Section Three details the system architecture and each component that makes up the system. The simulation environment application used in VLAD is discussed in Section Four. Section Five gives an overview of the interactive devices used. Section Six describes the model implementation at an object level. Section Seven gives an overview of the after-action review. Section Eight details the system controller, which is responsible for synchronizing and oversee the entire system. A conclusion is provided in Section Nine.

II. BACKGROUND

Traditional medical training focuses on technical skills by utilizing procedural instruction provided by educational studies or expert practitioners. Non-technical skills such as decision making, leadership, and situational awareness are difficult to assess during such training [2]. Without teaching methods that target these non-technical skills, many weaknesses are not discovered until workplace exposure. The apprenticeship method of instruction provides students experience by allowing them to observe and eventually involve themselves in the medical procedure. Medical simulation training allows for more effective evaluations of non-technical skills. EVMS currently uses a phantom based interactive birthing simulator known as SimMom, which is developed by Laerdal. Simulation-based training allows students the repeated training of various medical specialties without the consequence of harming a patient, and also allowing instructors to introduce complications rarely seen in the real world.

The VLAD project extends the physical simulation by projecting a virtual environment around the Manikin. The addition of this projection improves the training of the student by providing an immersive environment to use the Manikin in. VLAD is extensible and is able to be used on other medical procedures. There are non-interactive virtual devices representing real equipment such as a tocometer. The student interacts in the simulation by either performing an action on the Manikin or by interacting with an interactive virtual devices such as an IV pump. The phantom simulation enables a realistic physical model of the patient and fetus; however the experience of a true delivery can be improved through the addition of an immersive environment. VLAD attempts to expand on phantom simulations by adding a virtual environment to immerse the student in a realistic situation while also teaching technical and non-technical skills.

III. SYSTEM ARCHITECTURE

The system architecture for the VLAD Simulation Application was designed around five main structures that each enhance the training environment. These five objects are the Simulation Environment Application, the Interactive Devices, the System Model, the After-Action Review structure and the System Controller. The system architecture is shown in Fig. 1.

Each training scenario has a unique system configuration. The flow of events for scenarios, which is represented by the

System Model, is reconfigurable to facilitate alternate training objectives. The simulated environment is displayed on a four projection screen system which surrounds the Manikin. The Simulation Environment Application drives this projection system and is likewise modifiable, allowing for different environment models to be loaded. Training is also enhanced by the addition of the IVD which represent real-world medical equipment such as IV pumps and cardiac monitors. The System Controller ties the other constructs together and synchronizes the entire system, driving the simulation and facilitating communications. Taken together, these objects create an immersive training environment that enhances the already existing instructional technology by adding additional fidelity and immersion.

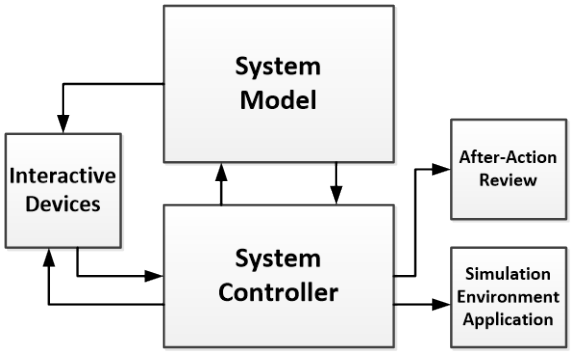


Fig. 1. System Diagram for the Simulation Application

A. Simulation Environment Application

The Simulation Environment Application (SEA) is responsible for displaying an environment model, maintaining Non-Interactive Virtual Device (NVD) statuses and playing audio. It communicates with other pieces of the VLAD system through a network interface. It is a stand-alone application that can be run on a separate computer to help alleviate computational strain.

B. Interactive Devices

The components that are designated Interactive Devices are the Interactive Virtual Devices (IVD), the Manikin, and the Instructor Tablet. These components drive the simulation by receiving and interpreting user input. Interactive Devices communicate with the rest of the VLAD system via a wireless network interface. Data is sent across this network in the JavaScript Object Notation (JSON) format. This provides a simple and fast way to encapsulate data for transfer and to un-package it upon receipt.

C. System Model

The System Model contains the state diagram representation of the scenario. As the scenario progresses, the state is updated when a transition event notification is received from the System Controller. When a state change occurs within the System Model, the controller is informed of a new set of conditions to monitor for future state changes. The System Model also triggers any updates for the IVDs, NVDs and Instructor Tablet.

D. After-Action Review

The After-Action Review (AAR) System has two primary functions during a simulation. The main function of the system is the logging of events as they are executed by the system controller. A log must be compiled in order for event tracing during AAR. The secondary function is to control the operation of the recording equipment.

E. System Controller

The System Controller is responsible for controlling and synchronizing the entire system. It contains two main software components: the Event Handler and the Simulation Executive. The Event Handler is the component that determines when the simulation transitions from the current state. It does this by monitoring the information received from the input devices, updating the current status of the system, and then checking the status against a set of applicable conditions for a set of state changes. When the Event Handler acknowledges full compliance to a set of conditions for a state change, it creates a transition event that is passed to the Simulation Executive.

The Simulation Executive receives events and execution times from each of the system components. When the execution time of a received event matches the current simulation time, the Simulation Executive executes that event. During execution of an event, the Simulation Executive notifies the AAR System and System Model of any information updates relevant for their functionality. The Simulation Executive also communicates with the SEA, informing it of any status changes to NVDs.

IV. Simulation Environment Application

The Simulation Environment Application operates as a separate application from the Simulation Application and it is in full control of the visualization and audio of the simulation environment. The visualization of SEA is implemented using the light-weight, open-source, 3D-graphics engine, OGRE. The structure of the visualization application is comprised of six primary objects: Message Handler, Simulation Executive, System Handler, Device Handler, Visualization Interface, and Visualization Engine. A system diagram for the SEA is shown in Fig. 2.

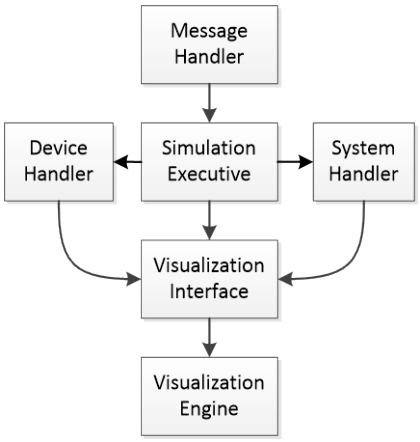


Fig. 2. Simulation Environment Application System Diagram

A. Message Handler

The Message Handler object is responsible for receiving a JSON string from the Simulation Application. During scenario initialization, the Message Handler receives the initialization messages which contain the required information for system setup. Throughout the rest of the simulation, messages contain either user command information or device update information that is used to instantiate a specific command pattern for a method on either the System Handler or the Device Handler, which is executed by the Simulation Executive.

B. Simulation Executive

The Simulation Executive object in the SEA is used to execute command patterns. This Simulation Executive is light-weight and simple in that it executes commands for only two other objects and executes in FIFO order. As messages are received and translated by the Message Handler, the executive receives them and stores them in an event list. The executive then processes all current events in order. Lastly, an update command is called on the Visualization Interface.

C. System Controller

The System Handler object is responsible for maintaining the current SEA state. It controls the simulation run state (paused, running, terminated), volume levels, current environment, scenario state, and audio library. System commands are routed and executed through the System Handler. When the Simulation Executive executes a command pattern created from a command message, the System Handler maintains the state that the command impacts and calls the control method on the Visualization Interface for that command.

D. Device Handler

The Device Handler object creates and maintains all NVD placed within the environment. Upon scenario initialization, the Message Handler receives messages containing device setup information which is then used to name a NVD and assign certain attributes through the Device Handler. Throughout the simulation, as the Simulation Executive executes command patterns created from device update messages, update methods are called on the Device Handler. The handler then updates the internal device attributes and passes along these updates to the Visualization Engine through device update methods on the Visualization Interface.

E. Visualization Interface

The Visualization Interface object provides a specific place in the application to route all visualization commands. By implementing the interface it becomes easier to utilize different visualization engines. The Visualization Interface contains methods to act as a translation service between the application handlers and the Visualization Engine.

F. Visualization Engine

The Visualization Engine object is implemented using OGRE and directly controls the viewports, virtual environment, and cameras. The engine also renders and controls the user interface portion for system setup. Using the setup interface, the user is able to select the projectors, resolution, anti-aliasing, rendering engine, and display positioning. The Visualization Engine is the sink of all commands that are sent to the SEA, and all commands are called through the Visualization Interface. Direct calls are used to alter system variables caused by commands and device updates; however, no changes are visible to the user until the Simulation Executive initiates an update to the Visualization Engine.

V. Interactive Devices

Interactive Devices provide an additional avenue for the student or instructor to interface with the simulation. Besides the Manikin, all other interactive devices are developed to run on Google Nexus 7 tablets that are running Android OS. The tablets display a Graphical User Interface (GUI) that the user is able to use to provide input to the simulation. The GUI in Android is implemented with XML layout files and Java source code files. The XML layout files are used to set the initial layout of the GUI, and any alterations to the GUI that occur due to button presses are handled in the source code. The XML layout files have been specifically laid out to fit the screen size of the Nexus 7. The XML files and Java files are compiled into an executable application that is run on the Nexus 7 tablet. This application takes input from the user and sends it across the wireless network to the Event Handler object. This allows the user's input to be logged on the Simulation Executive as an event.

A. IV Pump

The IV pump is an IVD that allows the user to change the flow rate of the IV fluids during the simulation. The GUI and the functionality of the IV pump are taken directly from the user's manual of the Abbot Hospira Plum A+ Infusion Pump [1]. The IV pump has two separate pumps that can be run simultaneously, which are labeled Pump A and Pump B. The default screen displayed while the pumps are running shows the rate of infusion in mL/hr and the volume infused in mL. The top of the screen reads PUMPING or STOPPED depending on whether the pump is running or not. This screen also has four softkey labels at the bottom that are coupled with softkey selectors as shown in Fig. 3. The middle two buttons are implemented in the application to include the functionality of the actual pump. If the user selects either A or B using the softkey selector button, then the screen changes to the selected pumps programming screen.

The programming screen is shown in Fig. 3. This screen allows the user to change the rate of infusion and volume to be infused for the selected pump. The user can scroll up and down between the rate and the volume to be infused using the up and down select button. As the values of the rate and volume to be infused are entered the duration of the treatment is automatically computed in hours and minutes if the duration takes longer than one hour or minutes and seconds if the duration takes less than one hour. After the intended rate and volume to be infused are entered into the virtual pump the

user selects start and the screen is automatically changed to the running screen. The rate and volume to be infused is then sent to the event handler across the wireless network.

Fig. 3. Developed IV Pump Interactive Device

B. *Laerdal Manikin*

The Laerdal Manikin is an important component that drives the scenario. The Manikin represents a patient during the training simulation. [5] The VLAD system uses the Laerdal Software Development Kit (SDK) in order to communicate and pull information from the Manikin. The connection to the Manikin is first done on the SimMan 3G server application. The SimMan 3G acts as the virtual Manikin and is used as a testing feature. The connection is written in C# and follows the SDK example shown in SimpleWindowsForm demo. The SimpleWindowsForm demo shows how to connect to the SimMan 3G server. A connection is made by matching a string server name to the Uniform Reference Identifier (URI) string. The URI is then used to attempt connection.

The WinFormsDemo project provided by the SDK demonstrates the manipulation of parameters. The SDK provides two types of parameters *App-Parameters* and *Model-Parameters*. Model-Parameters are related to the physiological simulation of the human body. App-Parameters deal with variables related to the Server such as pause or play. The first step is to declare an interface to the parameter. Iparameters are interfaces to the Manikin's attributes, which can be doubles, integers or Boolean. The Iparameter is initialized by matching a string to a value. Then a PropertyChangedEventHandler is set to notify changes to the Iparameter. Next, a function to return the Iparameter's value is constructed. A value such as heart rate would be a double. Parameters associated with pause and play are Boolean.

Communication between the Manikin and the VLAD system uses a C# client protocol in order to communicate with the operation computer. The VLAD Manikin software uses a C# socket in order to communicate with the operation computer. The operation computer subscribes to the Manikin in order to pull attribute values. The JSON format is used as communication protocol between the C# and C++ applications.

VI. System Model

The System Model is represented with a finite-state machine and is made up of a set of states, and each state has a set of transitions. In order for a transition to take place, there are two types of conditions that must be met. These two types of conditions are physiological conditions and instructor checklist conditions. The model is then implemented using C++ programming language.

A. *Model Implementation*

The Model is implemented using object-oriented program design. The first type of condition is implemented with a physiological condition object. Physiological conditions are used to monitor changes that happen to the manikin. An example of such a change is heart rate. If there was a certain heart rate range that needed to be met before a transition could occur, then this heart rate range would be a physiological condition for said transition. The Manikin's heart rate would need to be within the specified range in order for the condition to be met. Data about pertinent physiological attributes that need to be monitored during the training session is pulled from the Manikin at set intervals and used to evaluate the condition.

The second type of condition is implemented with the instructor condition object. Instructor checklist conditions are used to monitor specific actions that the student should be performing during the training simulation. An example of such a condition is student correctly identifies fetal head position [3,4]. This instructor checklist condition is checked either Yes or No depending on if the student correctly performs the task. As long as one of the boxes is checked, the condition is considered met. All of the instructor checklist conditions are displayed on the instructor tablet. As the scenario progresses each condition is checked off on the instructor tablet. When a condition is checked, the string is sent to the event handler and evaluated to see if it has been met.

The transition object is used to implement the state change that occurs in the clinical model once all conditions are met. In order to accomplish this, the object has three separate lists that hold physiological conditions, instructor checklist conditions, and any audio or visual events the scenario developer wants to occur during the next state.

The state object is used to implement the different states in the clinical model. Each state in the clinical model has a list of transitions that can occur.

The final step in Model implementation is taking all of the pieces and putting them together to make the actual model. The model object contains a list of states and a pointer to the current state of the Model. The command pattern used to represent state transitions is a transition event command pattern. The command pattern is created and scheduled on the

simulation executive at the current simulation time. When the command pattern is executed the transition event method of the model object is called. The first step in this method is to create a command pattern for each accessory event associated with the transition being executed. The transition object's accessory event list is iterated through and each accessory event is used to create either an audio or visual command pattern depending on the accessory event's type. The newly created command pattern is given the name of the file that needs to be executed and scheduled at the current simulation time plus the time increment noted by the accessory event. After all of the accessory events for the transition have been handled, the clinical model changes to the next state. Once the model changes state, it updates the event handler by sending it the list of transitions that can occur in the new state.

VII. After-Action Review

The After-Action Review System logs events during a simulation run. When the system is initialized to start a scenario, the student name, instructor name, date and time are recorded. When an event is executed, a snapshot of the system is taken. The snapshot records the event name and the time of execution. The current state, previous state, and next possible transitions associated with the executed event are also recorded. Finally, instructor learning conditions and important clinical attribute information are obtained. The information is generated as a text file that is read in the AAR GUI to set up the event history window. A video event is recorded just like any other event except an extra string is added to represent the video name. The video name is matched to the correct video file type and played by pressing the hyperlink available when selecting the video event.

The recorded events are shown in the AAR GUI in an event history window. The AAR GUI is used to review events one at a time. The event history window provides three separate windows. The state information is shown in one window, while the clinical attributes are shown in another. The last window associated with an event is the instructor checklist conditions, which shows a Yes or No answer along with the name and time of the learning event. An instructor comment box is included for any additional notes. The instructor is able to sit with the student and show which events the student needs to improve on.

VIII. System Controller

This section contains a description of the purpose, functionality, and communication architecture of the software constructs that make up the System Controller. Four software constructs make up the System Controller, the Simulation Executive, the Event Handler, the Run Time Interface (RTI) and the Network Handlers. The architecture of the system controller is shown in Fig. 4.

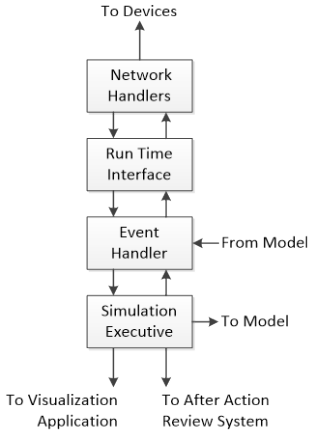


Fig. 4. Diagram of the System Controller Architecture

A. Simulation Executive

The Simulation Executive object maintains the list of pending events and executes them in time-stamped order. The Simulation Executive contains the main controlling loop called *RunSimulation()*. The Simulation Executive It is time stepped in that it waits for short periods of time before checking if there is an event at the current simulation time. The time step length is two seconds. The Simulation Executive also contains several methods for scheduling events. These *Schedule()* methods are called by other simulation objects, most notably the Event Handler. They create structures that contain all necessary information to execute events and store those structures in the pending event list. The three variants are: *ScheduleNow*(), which schedules the event at the current simulation time so it is executed during the next time step; *ScheduleDeltaT()*, which schedules the event DeltaT seconds in the future; and *Schedule()* which schedules an event for a specific simulation time.

Another notable aspect of the Simulation Executive's functionality is the play, pause, and stop methods. These methods are used by the instructor to control the playback of the simulation. The pause method allows for the training session to be stopped and resumed later. The *Stop()* method immediately ends the simulation and starts the preparation of the AAR.

B. Event Handler

The Event Handler object is responsible for interpreting output from the IVDs and the Manikin as well as scheduling events based on the current attribute values. It contains a list of the current possible transitions. When a new message is received by the Event Handler, a receiver thread is started. This receiver thread processes the message, responding to commands and updating attribute values. Conditions associated with any incoming attribute updates are checked along with the related transitions. At the end of the loop, a maximum of one active transition is scheduled and the current transition list is purged. In the very unlikely event that two transitions become active simultaneously, only the first transition checked will be scheduled.

C. Run Time Interface

The Run Time Interface structure manages the communication between simulation objects by routing messages to their intended recipients. It maintains lists of current subscriptions, attribute ownership, and interface references. When the RTI receives a new message, it kicks off a new receiver thread, assuming that one has not already been created. The receiver thread splits the incoming JSON object into several objects that are then sent to the intended recipients. The number and content of the JSON objects varies depending on the content of the received message.

D. Network Handlers

The Network Handler utilizes a socket based network architecture is created in order to facilitate communication between the SEA, IVDs and the System Controller. These interfaces are stored on the RTI. The generic Network Handler object is created to provide a common structure for all of the network objects, independent of platform or purpose. The implementation of this object varies slightly depending on whether it is acting as a server or a client; however, the basic functionality remains the same.

Each Network Handler object maintains two threads, one for input and one for output. The input thread utilizes a blocking receive function to detect when new messages have been sent across the network. In the event that the other end of the connection disconnects, the receive function terminates with a handled exception. When a proper message is received, it is processed into a JSON object which is then passed to the connected device or simulation object by calling the *ReceiveMessage()* method. The output thread is started when the *SendMessage()* method is called. It parses the JSON object that is being sent into a string and writes that string to the output stream. In the event that the output thread is already running when the *SendMessage()* method is called, the message is simply added to a queue and the output thread will process it before exiting that instance.

IX. Conclusion

The VLAD project represents a fusion between physical and virtual phantom based training. The synchronization between Laerdal's Manikin, the accurate IVDs, and the realistic environment model provides a unique level of environmental immersion. Additionally, the robust AAR System allows instructors to evaluate students and assess whether training goals have been met.

One of the other main benefits of the VLAD system is its extensibility. Though it was originally intended for obstetrics training, the software that was developed can easily be expanded to encompass a wide range of training domains. Additional IVDs, NVDs, and environment models can be developed and used to facilitate training in other clinical environments.

Acknowledgments

We would like to thank our instructors at Old Dominion University, Dr. Roland Mielke and Dr. Michel Audette, for mentoring us in this project; as well as our teaching assistant, Andy Ren. We would also like to thank the following individuals for serving as our clients: Geoffrey Miller and Andrew Cross of EVMS, Taryn Cupper and Bob Armstrong of the EVMS National Center for Collaboration in Medical Modeling and Simulation, and Curtiss Murphy of Alion Science and Technology.

References

[1] Hospira, Inc. System Operating Manual For Infusion Systems with v11.6 Software.

[2] Shah, Anand. Carter, Thomas. Kuwani, Thungo. Roger, Sharpe. (2013) Simulation to Develop Tomorrow's Medical Registrar. *Clinical Teacher, March 2013, Vol. 10, Issue 1*, pp. 42-46. 5p.

[3] Simmons, Carol L. (2012) Normal Vaginal Delivery Scenario Set Consolidated Instructor Manual. Available: www.healthcaresimulationsc.com/simstore/

[4] Patterson, Dale A., Winslow, Marguerite, and Matus, Coral D. (2008, August). Spontaneous Vaginal Delivery. *American Family Physician*, Volume(78), pp. 336-341. Available: www.aafp.org/afp

[5] Laerdal Simulator SDK (Version 1.0.0.294) Software Developer Kit (2010-10-11). Laerdal Medical AS. http://www.laerdal.com/us/SimMom

Occurrences of Delayed After-Depolarizations and Triggered Activity in RNC Model

Sharmin Sultana, Tanweer Rashid, Christian Zemlin

***Abstract* — Delayed after-depolarizations (DADs) are one of the suspected reasons of atrial fibrilation. They are mainly generated from cellular [Ca^{2+}] overload. Spontaneous [Ca^{2+}] release from the sarcoplasmic reticulum (SR) activates the transient inward current I_{ti}. This current is chiefly composed of several ionic currents such as the Sodium-Calcium exchange current I_{NaCa}, the Calcium Activated Chloride Conductance $I_{Cl(Ca)}$, and the non-specific Calcium current I_{ns-Ca}, and these contributes to DADs. This paper is a simulation study on the basic RNC atrial model to show occurrences of DADs under certain conditions.**

***Index Terms*—Delayed after-depolarization (DAD), Early after-depolarization (EAD), Sarcolplasmic Recticulam (SR), Calcium induced Calcium Release (CICR), Calsequestrin (CSQN).**

I. Introduction

After-depolarizations are oscillations in membrane voltage which occur before or after the completion of an action potential and depend on the preceding transmembrane activities. After-depolarizations might occur before or after the completion of the repolarization phase, and are called Early After-Depolarizations (EADs) or Delayed After-Depolarizations (DADs), respectively (Figure 1 and Figure 2). When EADs or DADs reach the activation threshold, they might generate one or a series of spontaneous action potentials which are called triggered activities or rhythmic activities.

Atrial fibrillation (AF) is one of the most common and sometimes dangerous types of arrhythmia. AF is an irregular pattern of heartbeats. During atrial fibrillation, the atria of the heart will start to fibrillate, and this might result in heart failure or brain stroke. One of the suspected mechanisms behind atrial fibrillations is Triggered Activities which might be induced by DADs [3, 5].

In healthy adult hearts, the electrical signal begins in the sinoatrial node (SA node) and then travels through the atria, enabling it to contract and pump blood into the ventricle. In AF, the electrical signal originates near the roots of the pulmonary vein instead of SA node. The signal then travels through the atria in an unsynchronized fashion and causes the atria to start to fibrillate. The heart starts to beat faster and in an irregular manner. As a result, blood cannot be pumped out properly from the heart and blood clots might form in the heart causing heart failure or brain stroke.

The special structure of the pulmonary vein might initiate DADs, which might eventually be strong enough to cause triggered activities [1]. In this paper we explored the conditions under which DADs and Triggered Activities might occur and then present simulation results of an existing atrial model (RNC model introduced in [2]) which was modified to be able to generate DADs and Triggered Activities.

II. Background and Literature Review

A. Action potential and its Phases

Cardiac cells are highly specialized cells which perform the conduction of electrical impulses and mechanical contraction. As cardiac myocytes are excitable, they have the ability to respond to a stimulus which produces Action Potentials (APs).

Once a cardiac cell is electrically excited, it begins a sequence of actions divided into 5 standard phases, and then the electrical signal is propagated to the adjacent cell. This process continues through all the cells of heart [3].

- Phase 0: This is the depolarization phase. When the membrane voltage reaches the activation threshold, Na^+ channels are triggered to open, which results in a large but transient Na^+ current. This current makes the membrane potential more positive.
- Phase 1: In this phase, the Na^+ current is inactivated and produce a subsequent K^+ current, and thereby initiates the repolarization phase.
- Phase 2: L-type Ca^{2+} current has an important effect in Phase 2 which is also called the plateau phase. In phase 2, the L-type Calcium channels are opened and remain open to give the AP a sustained phase. The L-type Ca^{2+} current triggers Ca^{2+} release from the SR which results in

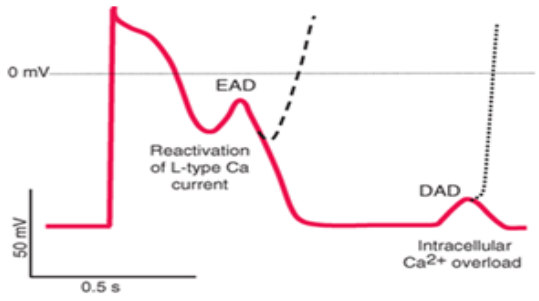


Figure 1: EADs and DADs in Action potential

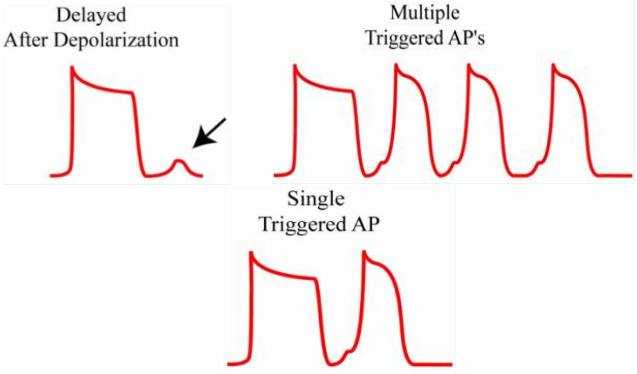


Figure 2: Generation of Triggered Activities from DADs. [http://www.fmhs.uaeu.ac.ae/wlammersteach/D.CVS/D.6.CVSPathophysiologie/D.5.3.ArrhythmiaA.html]

contractions of cardiac myocytes.

- Phase 3: This phase is called the rapid repolarization phase. During this phase, the L-type Ca^{2+} channels close and K^+ channels activate.
- Phase 4: This phase is the final repolarization phase where the cell reaches its resting potential. This is the phase where DADs are observed.

Figure 3 shows the phases of cardiac action potential and Figure 4 shows an atrial cell with all the ionic pumps, ionic channels and ionic concentrations.

A. Underlying cell Mechanism

During contraction, the $[Ca^{2+}]$ concentration needs to be balanced in all the intracellular compartments so that $[Ca^{2+}]$ cannot accumulate and impede contraction. The elevation of $[Ca^{2+}]$ ions in the intracellular spaces during a plateau phase causes diffusion of $[Ca^{2+}]$ ions into the SR through the uptake current called I_{up}. The SR has two compartments called Network Sarcoplasmic Reticulum (NSR) and Junctional Sarcoplasmic Reticulum (JSR). $[Ca^{2+}]$ ions move in between these two compartments through the translocation current I_{tr}. When the JSR has sufficient $[Ca^{2+}]$ concentration, it triggers spontaneous $[Ca^{2+}]$ release from the JSR through the release current I_{rel} which is known as the Calcium induced Calcium Release (CICR) process.

One of the ionic currents responsible for DADs and triggered activities is the transient inward current (I_{TI}) which is activated by spontaneous $[Ca^{2+}]$ release from the SR.

B. Related Works

A current contributing to DADs and triggered activities in mid-myocardial cells is the Calcium Activated Transient Inward Current I_{ti}-(I_{NaCa}) and the Calcium Activated Chloride current $I_{Cl(Ca)}$ [4]. Both of these currents activate during Calcium overload [4]. To prove the contribution of these current to DADs or triggered activities, one or the other current is blocked and tested to see whether they can produce DADs or triggered activity. By blocking the outward current $I_{Cl(Ca)}$, the exchange current I_{NaCa} is sufficient to produce triggered activities, shown in Figure 5. The outward current $I_{Cl(Ca)}$ was opposing the depolarizing influence of I_{NaCa}.

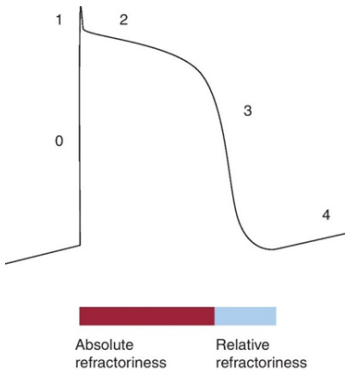


Figure 3: Action potential and its phases

By setting the Cl^- equilibrium potential as reversal potential and blocking I_{NaCa} the outward current $I_{Cl(Ca)}$ can still produce oscillation in membrane voltage [5]. Figure 5 shows that after inhibition of I_{NaCa}, there are still DADs and triggered activities.

An increase of the late sodium current (I_{Na}), which is a slowly inactivating depolarizing current, may also induce DADs and triggered activities [6]. An increase of I_{Na} eventually overloads the cell with $[Ca^{2+}]$, which in turn activates I_{TI}. ATX-II (a polypeptide toxin) was used to enhance I_{Na}. In purkinje and ventricular myocytes, 80% of I_{TI} consists of I_{NaCa} and the contributor of the remaining 20% is Cl^- current [7].

A complete mammalian ventricular action potential model was developed in [8], also called the Lou-Rudy (LR) model which implements the complete dynamics of intracellular $[Ca^{2+}]$ and also exhibits the behavior of the cell under $[Ca^{2+}]$ overload conditions. According to the LR model, when the cell is overloaded with $[Ca^{2+}]$, two types of release might occur from JSR. One is CICR, which is the normal release mechanism found in all mammalian cardiac cells. The other process is triggered when $[Ca^{2+}]$ concentration reaches a pre-defined threshold. When the release current activates, it eventually activates two other ionic currents which are the Na-Ca exchange current I_{NaCa} and the non-specific Ca^{2+} activated current $I_{ns,ca}$. Both of these currents contribute to the transient inward current ITI which in turn generates DADs and triggered activities.

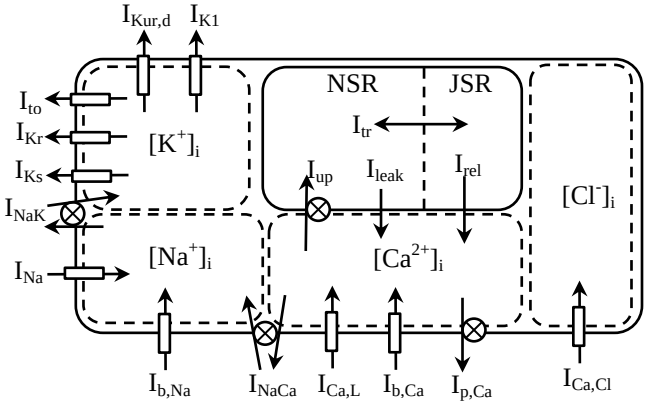


Figure 4: Schematic diagram of a canine atrial cell with all the ionic currents, concentrations and pumps. (Rafael J. Ramirez, 2000).

The LR model was built on mammalian ventricular cell and the author experimented under various [Ca^{2+}] overload conditions. If there is no [Ca^{2+}] overloading then the release mechanism is the CICR process, and there are very small DADs observed [8]. On the other hand, if overload conditions occur, then the release of [Ca^{2+}] generates large DADs and triggered activities.

The first mathematical model for canine atrial cell is the RNC Model, developed in [2]. It shows the basic mechanism of all the ionic currents during different phases of action potential. But the RNC model did not implement [Ca^{2+}] overload conditions, and therefore, could not be used to generate DADs and Triggered Activities.

III. RNC MODEL WITH LR OVERLOADED [CA^{2+}] RELEASE

The LR model implements two kinds of [Ca^{2+}] fluxes: The Calcium Induced Calcium Release (CICR) and the [Ca^{2+}] release of the JSR under [Ca^{2+}] overload conditions. CICR is the release of calcium under normal conditions, and is governed by the following equation:

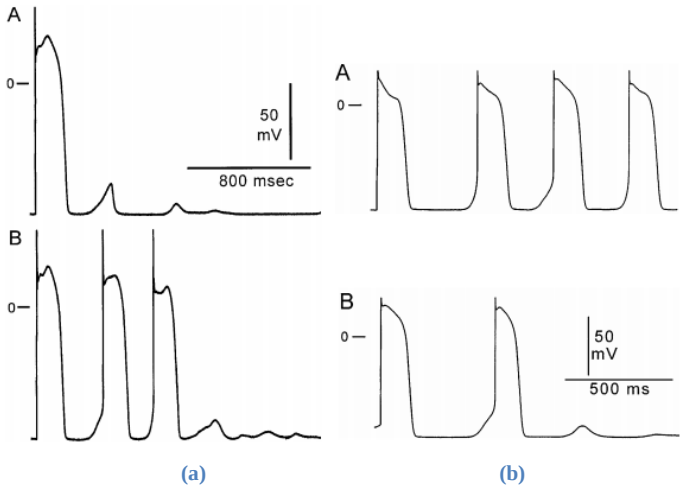
$$I_{rel} = G_{rel} \times \left([Ca^{2+}]_{JSR} - [Ca^{2+}]_i\right)$$

$$G_{rel} = \bar{G}_{rel} \times \frac{\Delta[Ca^{2+}]_{i,2} - \Delta[Ca^{2+}]_{i,th}}{K_{m,rel} + \Delta[Ca^{2+}]_{i,2} - \Delta[Ca^{2+}]_{i,th}} \times \left(1 - e^{-t/\tau_{on}}\right) \times e^{-t/\tau_{off}}$$

G_{rel} is a gating variable. The overloaded condition in the LR model occurs when the level of the calcium buffer calsequestrin (CSQN) exceeds a certain limit given by $CSQN_{th}$. When this happens, LR model redefines the gating variable G_{rel} as

$$G_{rel} = \bar{G}_{rel} \times f(t)$$

$$f(t) = \left(1 - e^{-t/\tau_{on}}\right) \times e^{-t/\tau_{off}}$$



(a) (b)

Figure 5: (a) Triggered activity in absence of $I_{Cl(Ca)}$: In A, the last stimulated beat is shown at left. In B, last stimulated beat is shown at left. (b) Effects of inhibiting I_{NaCa} on DADs and TA: In A, the last stimulated beat is shown at left. In B, the last stimulated beat is shown at left [4].

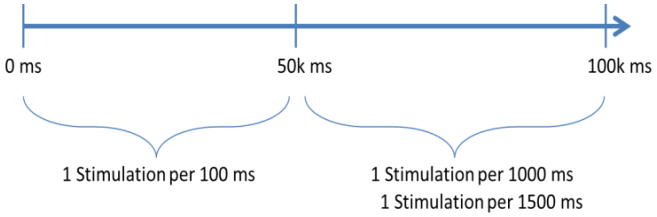


Figure 6: Stimulation phases for the experiment

The overloaded release current is supposed to have a very short duration, and the exponential function *f(t)* models the overloaded current as a relatively short-lived spike.

Our objective for this paper was to implement the overloaded [Ca^{2+}] release mechanism of the LR model into the RNC model. We implemented this concept of the LR model into the RNC model.

IV. SIMULATION RESULTS

A. Simulation Conditions

All simulations were run for 100 seconds (Figure 6). During the first 50 seconds, the model was stimulated at a rate of 1 stimulation per 100 ms (fast stimulation phase). During the last 50 seconds, the model was stimulated at a rate of 1 stimulation per 1000 ms or 1 stimulation per 1500 ms (slow stimulation phase). Fast stimulation causes calcium to build up in the cell, as shown in Figure 7.

The values of the following parameters were changed to see what effect they had on DAD and Triggered Activitiy generation.

- CSQNth
- JSR Volume
- τ_tr
- INaCa Sodium-Calcium exchanger.

B. Overloaded [Ca^{2+}] Release

Figure 8 shows a graph of the variables of interest between 50 seconds and 60 seconds. During the fast stimulation phase (i.e. for simulation time < 50 seconds), calcium had been building up in the intracellular spaces.

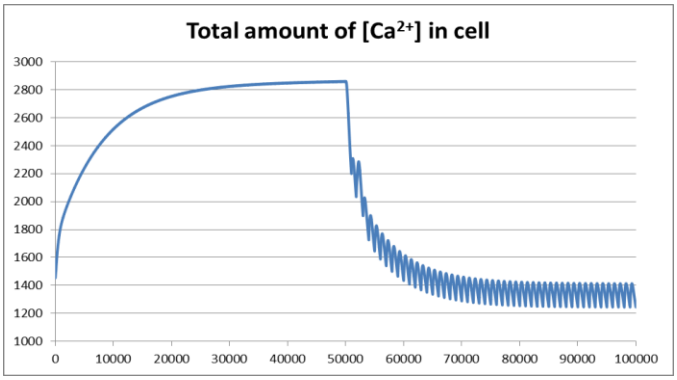


Figure 7: Buildup of calcium during the fast stimulation phase.

At simulation time = 50 seconds and onwards, the slow stimulation phase starts, and $[Ca^{2+}]$ ions are taken up into the NSR compartment of the SR. This causes a decrease of $[Ca^{2+}]$ ions in the intracellular volume. At the same time, the $[Ca^{2+}]$ levels in the JSR compartment rises because $[Ca^{2+}]$ ions get transferred from the NSR to the JSR. This is why at the beginning the $[Ca^{2+}]$ level in the NSR remains constant. However, afterwards, more and more $[Ca^{2+}]$ ions get transferred from the NSR to the JSR, and the $[Ca^{2+}]$ level in the NSR starts to drop. At simulation time of approximately 50.568 seconds (marked by the dotted red line on the left), the CSQN level reaches the threshold ($CSQN_{th}$ = 3.5 in this case), and this causes a release of $[Ca^{2+}]$ ions into the intracellular spaces (in other words, a DAD occurs). At this point, a sharp drop of $[Ca^{2+}]$ ions is seen in the JSR along with an increase of $[Ca^{2+}]$ ions in the intracellular spaces. This process goes on, until the overall $[Ca^{2+}]$ level in the NSR starts to decrease, and DADs occur less frequently.

C. Effect of Changing $CSQN_{th}$

Calsequestrin (CSQN) is a calcium-binding protein or calcium buffer of the sarcoplasmic reticulum. The LR model implemented the overloaded calcium release by defining a limit for CSQN. In the LR model this limit ($CSQN_{th}$) had a value of 0.7. However, under normal circumstances, the CSQN levels in the RNC model are always above 0.7, as shown Figure 9.

In Figure 9, the model is being stimulated at a rate of 1 stimulation per 100 ms for the first 100 seconds, and no stimulation from 100 seconds to 150 seconds. The CSQN level remains slightly above 6 at the beginning, then drops to around 1.5 during the fast stimulation phase, and then returns to the steady value of above 6 when no stimulation is applied.

The value for $CSQN_{th}$ was assumed to be somewhere between 3.0 and 5.0. Figure 10, Figure 11, Figure 12, Figure 13 and Figure 14 shows the effect of using different values for $CSQN_{th}$. As mentioned before, DAD generation is implemented by defining a value for $CSQN_{th}$. DADs are generated when the CSQN level exceeds this limit.

If the CSQN limit is too high, as in Figure 10, then DADs are not generated because the threshold limit is not being exceeded. However, in Figure 11, the threshold is 4.5, and that is being exceeded once just before 52 seconds in simulation time. In this case, the $CSQN_{th}$ value is relatively large, and so the DAD being generated has higher amplitude. In Figure 12, the threshold is 4.0, and this threshold is being exceeded thrice just before 51 seconds, 52 seconds and 54 seconds, respectively of simulation time. But the amplitude of the DADs is smaller than in Figure 11. In Figure 13 and Figure 14, the $CSQN_{th}$ value is 3.5 and 3.0, respectively, and the amplitudes are smaller.

Essentially, if the $CSQN_{th}$ is higher (but not exceeding a certain limit), then more $[Ca^{2+}]$ ions accumulate in the SR, and therefore DADs of larger amplitudes are observed. However, for high $CSQN_{th}$, fewer numbers of DADs are observed. Smaller values of $CSQN_{th}$ allow for lesser amount of $[Ca^{2+}]$ ions to accumulate in the SR, and so the DADs have smaller amplitudes. But for smaller $CSQN_{th}$, more numbers of DADs are observed.

D. Effect of Changing the Volume of JSR

The volume of the JSR is an important variable because this determines the amount of $[Ca^{2+}]$ ions that can be stored in the JSR. Figure 15, Figure 16, Figure 17, Figure 18 and Figure 19 show the effect of using different volumes of the JSR.

When the volume of the JSR is small, as in Figure 15, lesser amounts of $[Ca^{2+}]$ ions accumulate in the JSR. This is why the amplitude of the DADs is smaller. However, when the volume is larger, as in Figure 16, Figure 17 and Figure 18 larger amounts of $[Ca^{2+}]$ ions can accumulate in the SR, and DADs of larger amplitudes are observed. In Figure 18, the amplitude of the DAD is large enough to generate a triggered activity (TA). However, it has been observed that too large volumes do not allow for generation of DADs. In Figure 19, the volume is 150% of the original, and no DADs are observed.

In Figure 15, it can be seen that a smaller volume causes DADs to occur more frequently. The DADs occur less frequently when larger volumes are used. At this point, we do not have enough information to verify this conclusively, nor are we able to provide an explanation as to why DADs are more frequent for smaller volumes.

E. Effect of Changing τ_{tr}

τ_{tr} is the time constant of $[Ca^{2+}]$ transfer from NSR to JSR. As mentioned before, excess $[Ca^{2+}]$ ions are absorbed into the NSR from the intracellular spaces. These $[Ca^{2+}]$ ions are then transferred from the NSR to the JSR, and this transfer is governed by the following equation:

$$I_{tr} = \frac{[Ca^{2+}]_{NSR} - [Ca^{2+}]_{JSR}}{\tau_{tr}}$$

Essentially, if τ_{tr} is smaller, more $[Ca^{2+}]$ ions are transferred from NSR to JSR, and larger τ_{tr} means lesser $[Ca^{2+}]$ ions are transferred from NSR to JSR. The value of τ_{tr} for both LR and RNC model is 180. The following figure (Figure 20) shows the DADs for different values of τ_{tr} .

Figure 20 shows the DADs being generated for different values of τ_{tr}. As can be seen, for larger τ_{tr} the DADs are generated at later times. Smaller τ_{tr} values cause DADs to be generated earlier. In the case of τ_{tr} being 90 and 45 (purple and light blue curves, respectively), two DADs are being generated.

Smaller τ_{tr} means more $[Ca^{2+}]$ get transferred from NSR to JSR. So the $CSQN_{th}$ is being exceeded faster. Larger τ_{tr} means lesser $[Ca^{2+}]$ ions get transferred from NSR to JSR. So it takes much longer for the CSQN level to exceed $CSQN_{th}$, and so DADs are generated at later times.

F. Effect of Removing I_{NaCa} (Sodium-Calcium exchanger)

The Na-Ca exchanger removes calcium from the

intracellular spaces. So the intracellular calcium levels become smaller and smaller over time. If, for some reason, the Na-Ca exchanger is disabled, then the [Ca^{2+}] ions will not be removed, and DADs will be generated infinitely. This is shown in Figure 21.

Since the [Ca^{2+}] ions are not removed from the intracellular spaces, the [Ca^{2+}] ions are taken up by the NSR and subsequently get transferred to the JSR. When the CSQN threshold is exceeded, these [Ca^{2+}] ions are released into the intracellular spaces through DADs, and this cycle repeats.

V. Conclusion

In this paper, we implemented the DAD generating mechanism from the LR model into the RNC model. We experimented with different parameters ($CSQN_{th}$, τ_{tr}, JSR volume and Na-Ca exchanger) to see their effects on DADs. The RNC model is an atrial model, while the LR model was a ventricular model.

It has been observed that the value for $CSQN_{th}$ has an important effect on DADs. Higher values of $CSQN_{th}$ results in stronger DADs, however, too large values of $CSQN_{th}$ are detrimental to the generation of DADs. Another parameter studied was the transfer constant τ_{tr}. It has been observed that higher values of τ_{tr} causes DADs to be generated at later times, and smaller values of τ_{tr} cause the DADs to be generated earlier. If τ_{tr} is very small, then multiple DADs are generated between stimulations (the light blue curve in Figure 20). The volume of the JSR also affects the generation of DADs. A larger volume of the JSR can produce stronger DADs. However, if the volume is too large, then DADs are not produced. The last parameter examined was the Na-Ca exchange current. By blocking this current, weak DADs are generated throughout the simulation time. Under certain conditions, it has been observed that DADs can be sufficiently large enough to cause a Triggered Activity

VI. References

[1] P. A. B. Andrew L Wit, "Triggered Activity and Atrial Fibrillation," Elsevier, 2006.

[2] S. N. a. m. C. Rafael J. Ramirez, "Mathematical analysis of canine atrial action potentials:rate, regional factors, and electrical remodeling," Am J Physiol Heart Circ Physiol, 2000.

[3] F. E. M. a. B. P. B. Larraitz Gaztanaga, "Mechanisms of Cardiac Arrhythmias," revista espanola de cardiology , 2012.

[4] R. J. G. a. C. M. W. Andrew C. Zygmunt, "INaCa and ICl(Ca)contribute to isoproterenol-induced delayed afterdepolarizations in midmyocardial cells," AJP - Heart and Circulatory Physiology, 1998.

[5] K. N. T. K. S. N. Yuki Iwasaki, "Atrial Fibrillation Pathophysiology Implications for Management," American Heart Association, 2011.

[6] J. C. S. a. L. B. Yejia Song, "An increase of late sodiumcurrent induces delayed afterdepolarizations and sustained triggered activity in atrial myocytes.," AJP-Heart and Circulatory Physiology, 2007.

[7] P. Verkerk, P. Marieke W. Veldkamp, P. Lennart N. Bouman and P. Antoni C.G. van Ginneken, "Calcium-Activated Cl2 Current Contributes to Delayed Afterdepolarizations in Single Purkinje and Ventricular Myocytes," Circulation, 2000.

[8] C. L. a. Y. Rudy, "A dynamic model of the cardiac ventricular action potential. II.Afterdepolarizations, triggered activity, and potentiation," Circulation Research, 1994.

[9] G.-N. T. a. A. L. Wit, "Characteristics of a Transient Inward Current That Causes Delayed Afterdepolarizations in Atrial Cells of the Canine Coronary Sinus," J Mol Cell Cardiol, 1987.

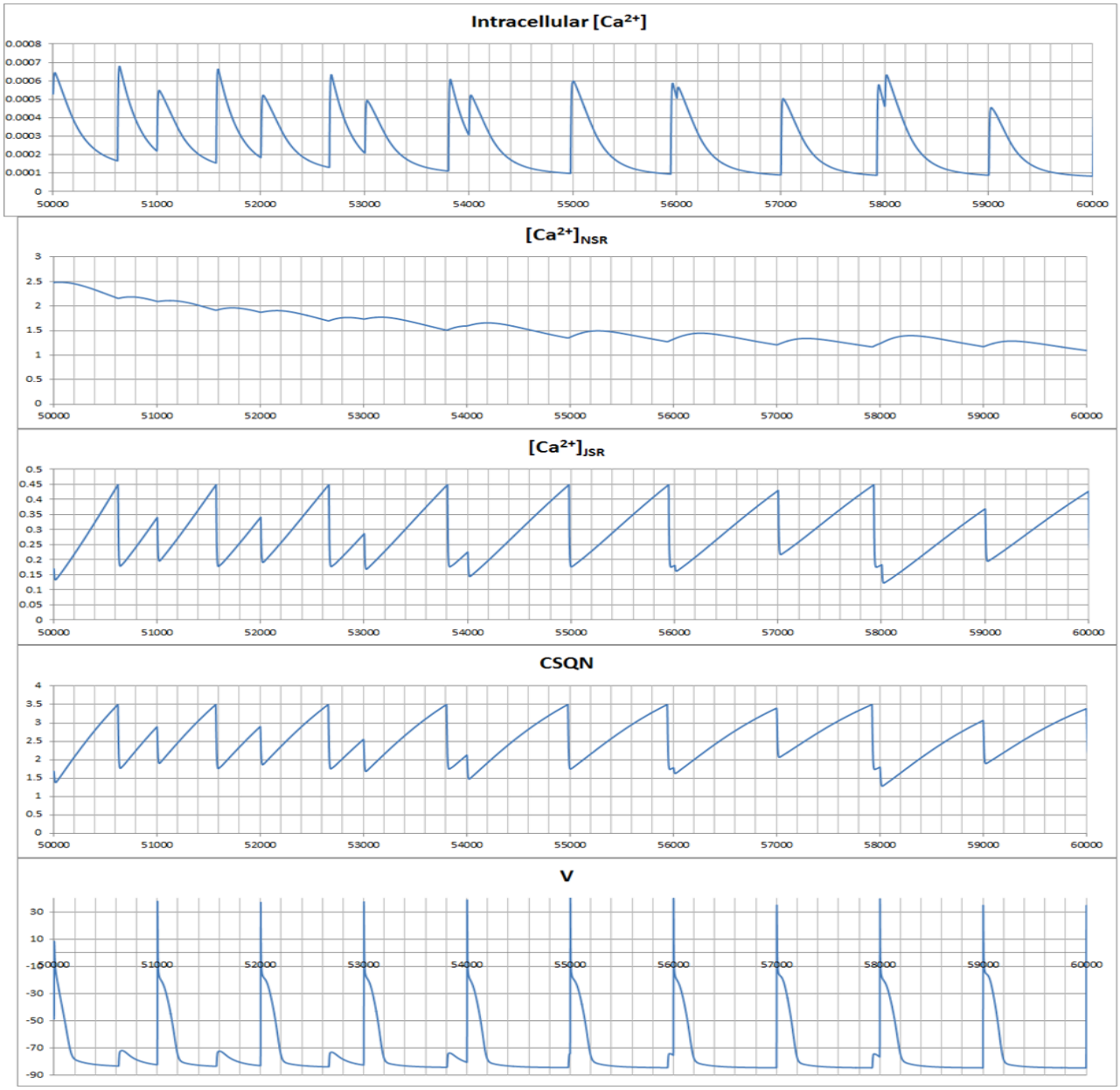


Figure 8: Values of different parameters in RNC model

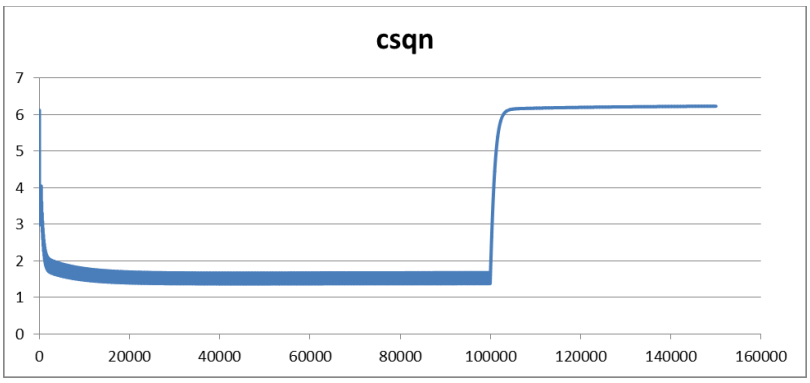


Figure 9: CSQN levels and membrane potential for CSQNth = 4.0

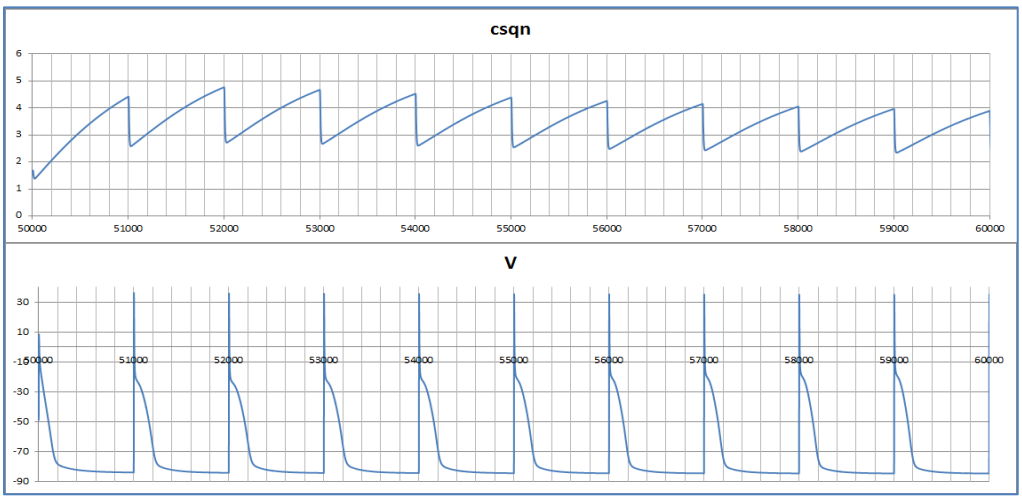


Figure 10: CSQN levels and membrane potential for CSQNth = 5.

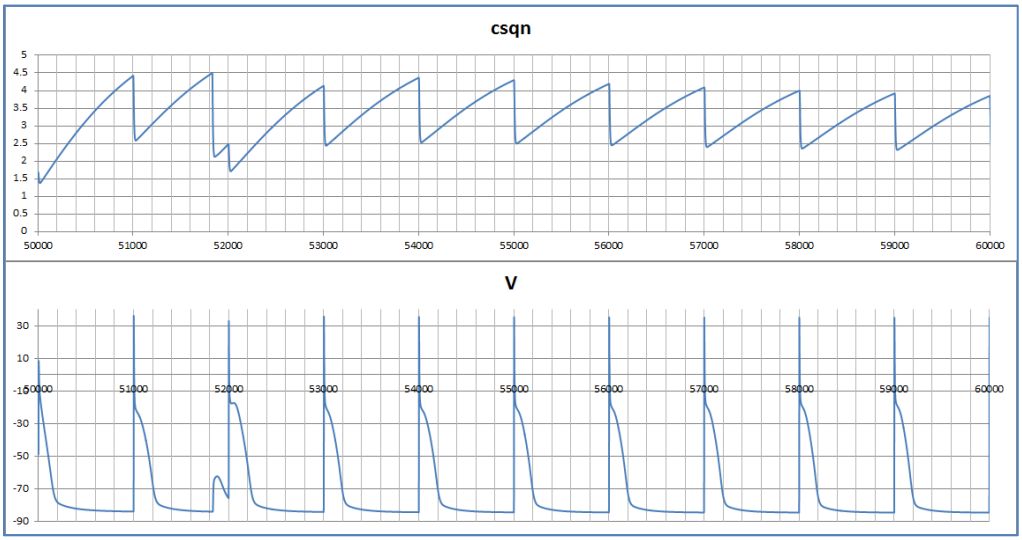


Figure 11: CSQN levels and membrane potential for CSQNth = 4.5

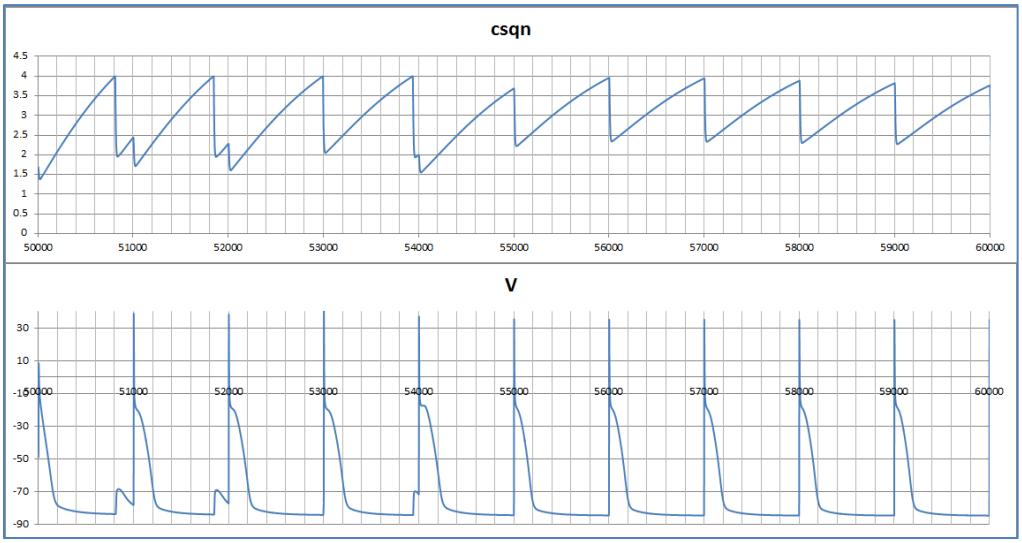


Figure 12: CSQN levels and membrane potential for CSQNth = 4.0

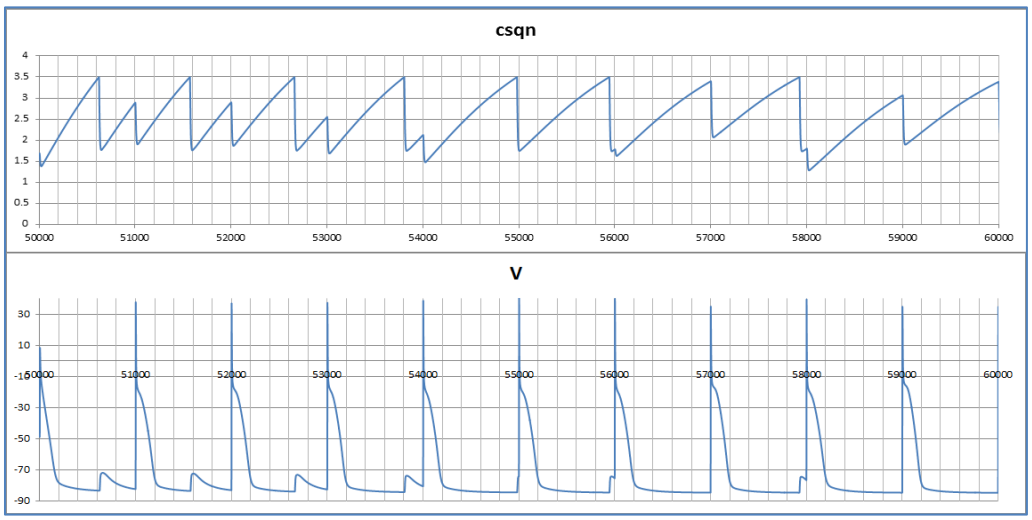


Figure 13: CSQN levels and membrane potential for $CSQN_{th}$ = 3.5

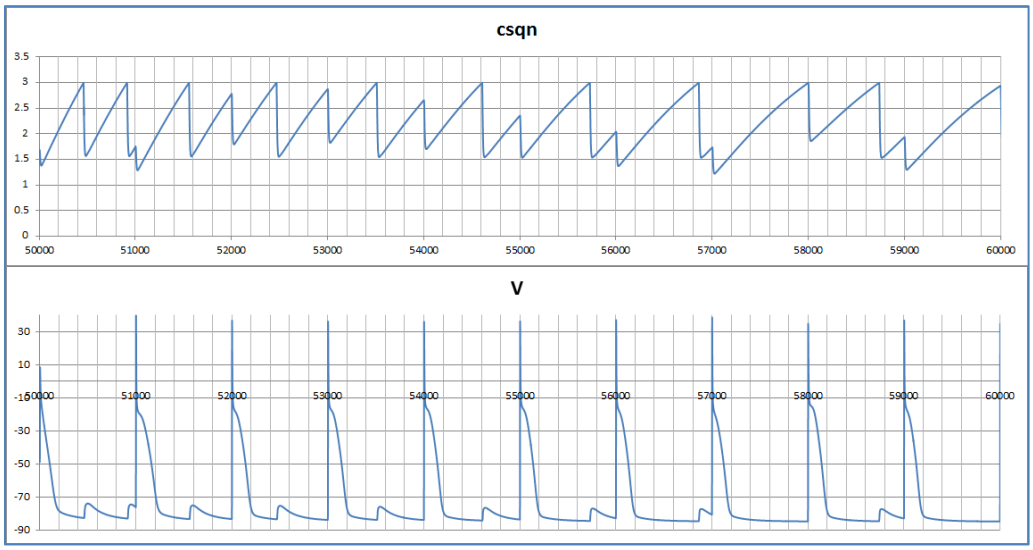


Figure 14: CSQN levels and membrane potential for $CSQN_{th}$ = 3.0

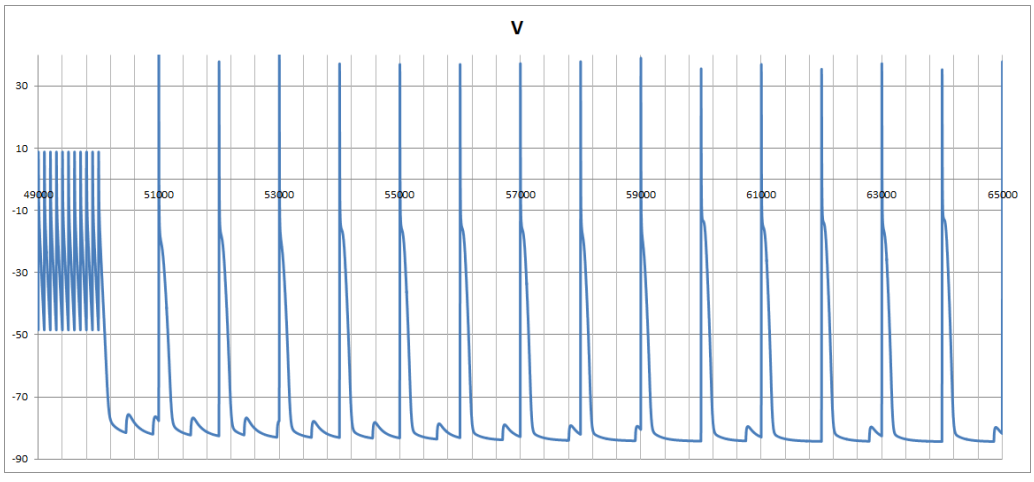


Figure 15: Volume of JSR is 50% of original.

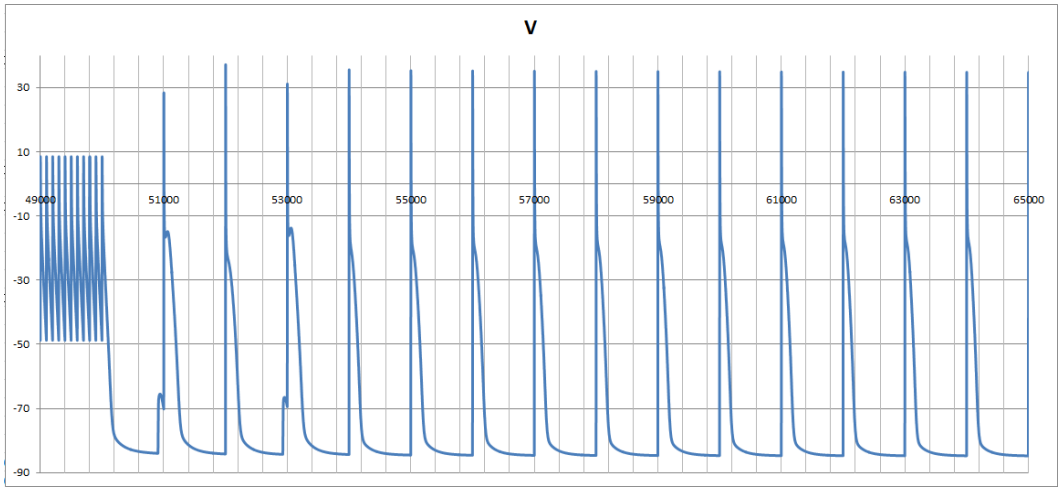


Figure 16: Volume of JSR is 110% of original.

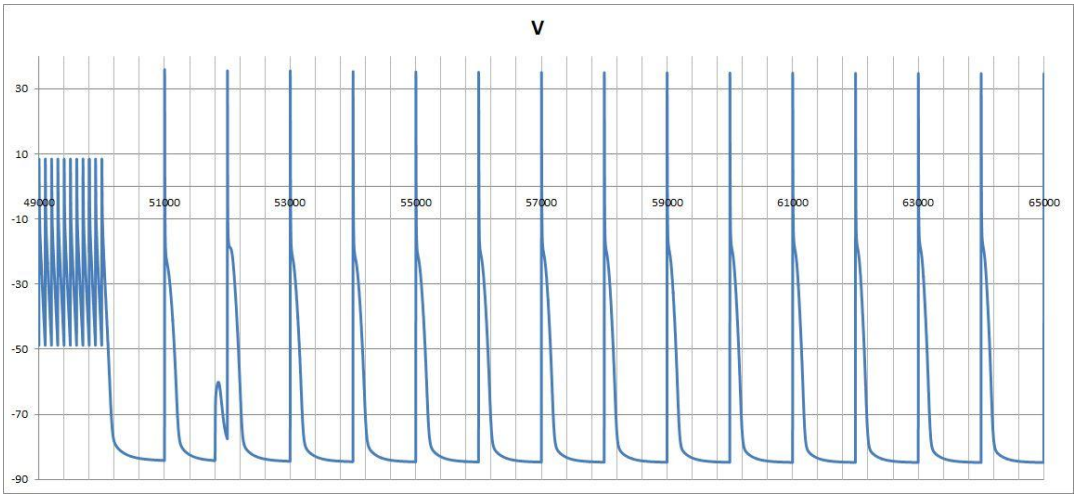


Figure 17: Volume of JSR is 120% of original.

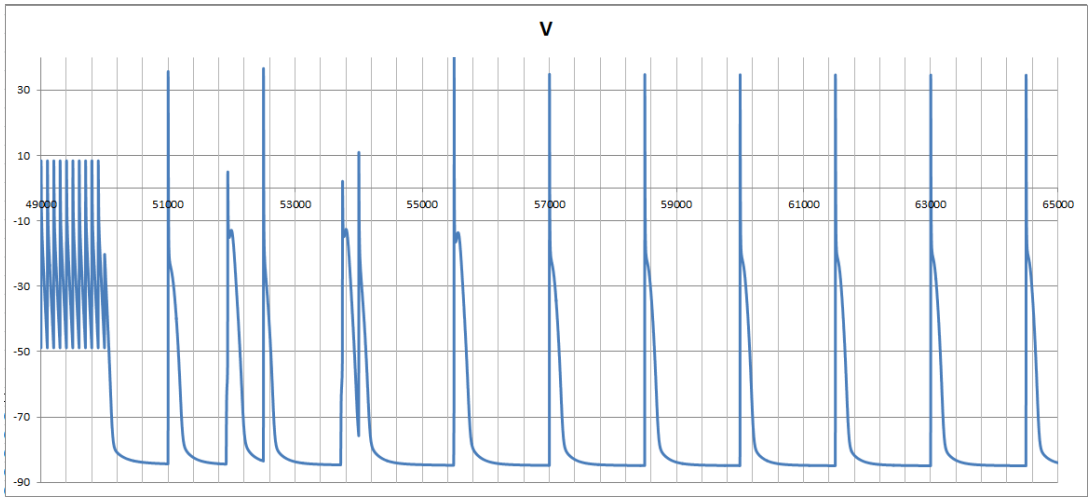


Figure 18: Volume of JSR is 130% of original.

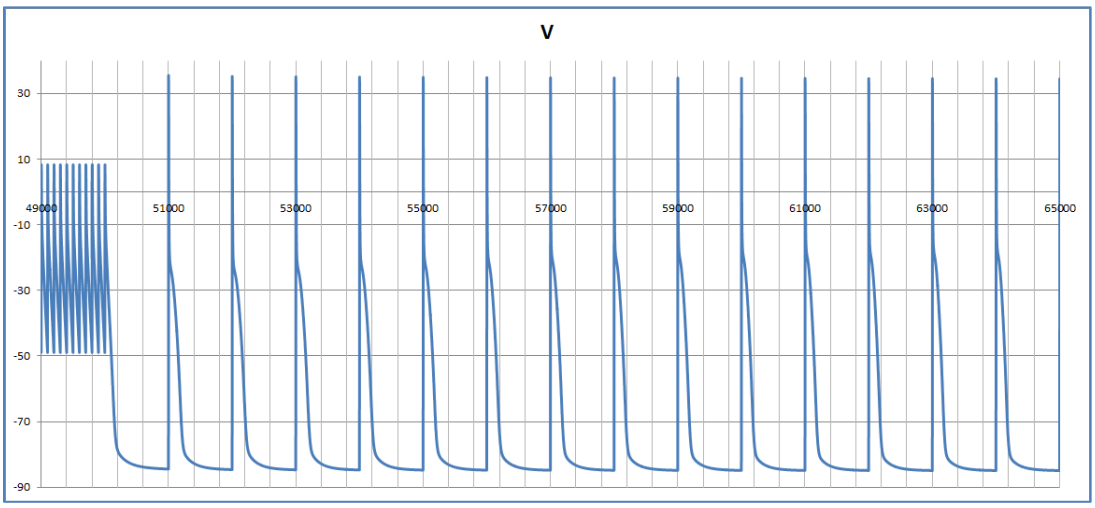


Figure 19: Volume of JSR is 150% of original.

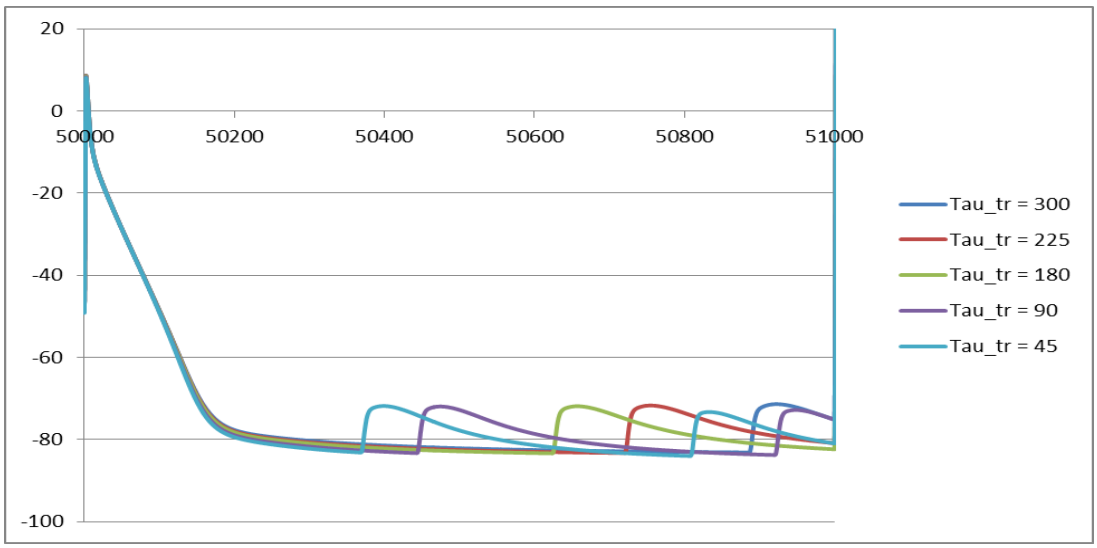


Figure 20: DADs for different values of τ_{tr}. The simulation time is between 50 seconds and 51 seconds, i.e. between two stimulations.

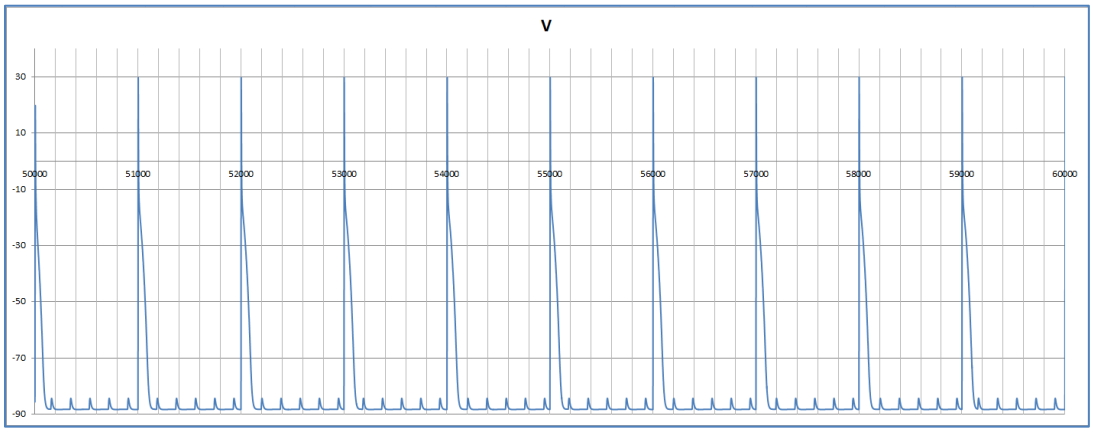


Figure 21: Effect of disabling the Na-Ca exchanger.

Modeling and Simulation of the Shape Recovery of Red Blood Cells

John Gounley, Yan Peng

I. Extended Abstract

The shape change of viscoelastic, fluid-filled capsules has received considerable attention from researchers in recent years. This attention has particularly centered on their application to red blood cells, which may be modeled in such a way. The passage of blood through small capillaries requires significant deformation by red blood cells, from the normal biconcave discoid to a bullet-like shape [8]. Upon reaching larger blood vessels, the red blood cells recover their normal shape. In blood diseases, such as sickle-cell anemia, the ability of red blood cells to deform and subsequently recover their shape is reduced. As a result, they may block capillaries and oxygen delivery may be adversely impacted [6]. To aid the development of treatments for such blood diseases, and to better understand the mechanical structure of red blood cells, it is important to study the manner in which their shape is deformed and recovered.

Extensive computational study has been made of the deformation of viscoelastic, fluid-filled capsules under shear flow. The roles of the membrane's elasticity, bending stiffness, and viscosity in the deformation process having been clarified, along with the effect of different fluid viscosities inside and outside of the capsule (e.g.,[4], [5], [7]). Conversely, investigations into the shape recovery of capsules from deformation have been largely limited to experimental and theoretical avenues, including optical tweezing and micropipette aspiration of red blood cells [2], [3]. These studies primarily aimed at measuring the time course of shape recovery and determining the dominant mechanisms by which it occurred. While relaxation from micropipette aspiration or optical tweezing may be a primarily solid mechanical process, as modeled in [2] and [3], Baskurt and Meiselman's work suggests that a similar characterization may be used to describe the more complex case of shape recovery of red blood cells from deformation by shear flow [1].

In this work, we introduce a 2D model of capsule deformation in, and shape recovery from, shear flow. A lattice Boltzmann method (LBM) is used to solve the fluid flow, while the immersed boundary method (IBM) is chosen to simulate the fluid-structure interaction. The structural model of the capsule includes shear elasticity, bending stiffness, and membrane viscosity. Additionally, the model allows for different fluid viscosities inside and outside of the capsule. We use the model to simulate shape recovery of the capsule after the abrupt stop of shear flow for various fluid and capsule parameters, for both circular and biconcave capsules. We find that while an exponential decay function $e^{-t/tc}$ fits the data very well for circular capsules, the recovery of tank-treading biconcave capsules is best described by a pair of recovery modes, which are characterized by different decay functions.

Our results differ from previous studies in their consideration of the entire shape recovery process. In the work of Evans and Hochmuth, Dao et al, and others, they consider recovery with respect to the principal stretch ratio λ and found $tc = \rho\frac{\eta_e}{E_s}$, for a constant ρ, membrane viscosity coefficient η_e, and shear elasticity modulus E_s. This is consistent our results for the intial recovery of biconcave capsules from tank-treading, but fails to accomodate the major shape changes which a tank-treading capsule must undergo later in its shape recovery. To capture these more substantial changes, we measure the shape recovery with respect to the Taylor deformation parameter D_{xy}.

Consequently, we posit a two-part recovery for tank-treading biconcave capsules. Illustrated in figures 1a and 1b, we consider the first mode between times A and B, and the latter mode between B and C. The first mode, focused on dissipating large intra-membrane forces quickly, involves λ nearly returning to its initial value and is due to the membrane's elasticity and viscosity. In particular, we find that the time constant for the first mode may be approximated as $tc = \rho GM$, for dimensionless shear rate G, and membrane viscosity ratio M; this equation is equivalent to that of Evans and Hochmuth. The second mode, unnecessary for the smaller deformations which occur in micropipette aspiration, dissipates remaining forces while membrane elements return to their initial positions and preferred curvatures. For this latter mode, we found that shear elasticity, the viscosity jump, and bending stiffness played significant roles in determining the time course of the shape recovery. We estimate the time constant of this latter mode as $tc = \rho\frac{GV}{E_b - \phi}$, for viscosity jump V, bending stiffness ratio E_b, and constant $\phi << 1$ representing the necessity of $E_b > 0$ for a non-circular capsule to recover its shape in 2D.

J. Gounley and Y. Peng are with the Department of Mathematics and Statistics at Old Dominion University.

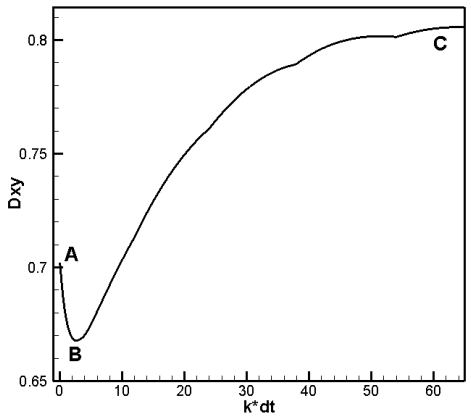
Acknowledgment

Support from the ODU Modeling and Simulation Initiative is gratefully acknowledged.

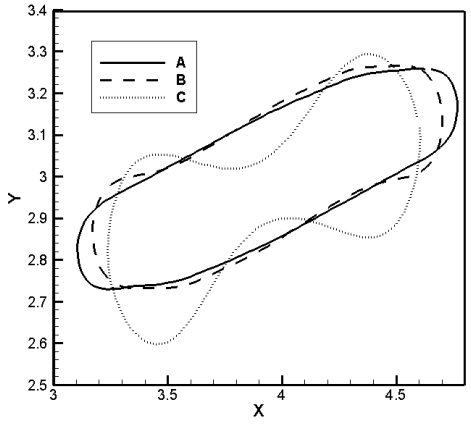
References

[1] O. Baskurt and H. Meiselman, *Determination of red blood cell shape recovery time constant in a couette system by the analysis of light reflectance and ektacytometry*, Biorheology, 33 (1996), pp. 487–501.

[2] M. Dao, C. Lim, and S. Suresh, *Mechanics of the human red blood cell deformed by optical tweezers*, Journal of the Mechanics and Physics of Solids, 51 (2003), pp. 2259–2280.



(a)



(b)

Fig. 1: Recovery of a tank-treading biconcave capsule in terms of (a) the Taylor deformation parameter D_{xy} and (b) the capsule shape. Shear flow is stopped at time A, the Taylor deformation parameter reaches a minimum at time B, and the capsule fully recovers its initial shape at time C.

[3] E. EVANS AND R. HOCHMUTH, *Membrane viscoelasticity*, Biophysical Journal, 16 (1976), pp. 1–11.
[4] C. POZRIKIDIS, *Numerical simulation of the flow-induced deformation of red blood cells*, Annals of Biomedical Engineering, 31 (2003), pp. 1194–1205.
[5] Y. SUI, Y. CHEW, P. ROY, X. CHEN, AND H. LOW, *Transient deformation of elastic capsules in shear flow: Effect of membrane bending stiffness*, Physical Review E, 75 (2007), pp. 066301–1 – 066310–10.
[6] S. USAMI, S. CHIEN, P. SCHOLTZ, AND J. BERTLES, *Effect of deoxygenation on blood rheology in sickle cell disease*, Microvascular Research, 9 (1975), pp. 324–334.
[7] J. ZHANG, *Effect of suspending viscosity on red blood cell dynamics and blood flows in microvessels*, Microcirculation, 18 (2011), pp. 562–573.
[8] J. ZHANG, P. JOHNSON, AND A. POPEL, *An immersed boundary lattice boltzmann approach to simulate deformable liquid capsules and its application to microscopic blood flows*, Physical Biology, 4 (2007), pp. 285–295.

A Parallel Marching Cubes Algorithm for Extracting Isosurface from Medical Images

Jing Xu and Andrey N. Chernikov

***Abstract*— The processing and understanding of medical images rely heavily on the ability to visualize the surfaces of the organs or the tissues represented in the images. This visualization capability is especially critical for three-dimensional images which are composed of a number of two-dimensional slices aligned and stacked on top of each other, such as the ones obtained with computed tomography, magnetic resonance, ultrasound, and other technologies. One of the most frequently used techniques for extracting the surfaces from medical images is the Marching Cubes algorithm. Given a required intensity threshold, this algorithm processes the volumetric pixels (voxels) in the image one at a time and constructs a number of triangles in the three dimensional space that approximate the isosurface corresponding to this threshold. The union of all triangles from all voxels yields a high accuracy polyhedral representation of the isosurface which is suitable for rendering on graphics hardware or for further processing. This extraction of isosurfaces is frequently used in interactive and/or iterative applications, and therefore the speed of extraction needs to be minimized. In this paper we present our parallelization of the Marching Cubes algorithm which achieves a nearly linear decrease in the processing time with respect to the number of used hardware cores..**

***Index Terms*—Medical image, Isosurfaces, Parallel, Marching Cubes**

I. Introduction

A number of applications which model complex physical and bio-medical phenomena require fast construction of very large surface meshes, marching cubes is a computer graphics algorithm for creating a polygonal mesh of an isosurface of a three-dimensional scalar field. It is used in a large number of applications for three-dimensional surface representation and visualization in such fields as finite element simulations and medical image computing. This algorithm is first published in the 1987 SIGGRAPH proceedings by Lorensen and Cline[1].

The previous published approaches identified a number of issues in paralleling the marching cubes algorithm. Since the load-balancing plays an important role in parallel performance, a number of authors focused on dynamic or static load balancing strategies. Gerstner T, Rumpf M[2] presented an efficient approach for tetrahedral grids recursively generated by bisection, and they parallelize it by giving each process a subtree which could be handled independently, dynamic load-balancing achieved by choosing the remaining unprocessed subtree of another process when process has already finished its whole subtree. This strategy prevents some processes from becoming idle while others are busy processing with large unvisited subtree. Chiang Y-J, Farias R, Silva C, Wei B[3] paid their attention to an out-of-core dynamic load-balancing scheme for cluster computers. They first constructed three files: meta-cells, which is clusters of cells partitioning the original dataset; BBIO tree and bounding-box files, which are used to index the meta-cells. When a client node is initially idle or finishes its current surface extraction on the meta-cell, it sends a job request to the host CPU. The host handles job requests from all other nodes using the BBIO tree and bounding-box file in the host disk to find all meta-cells. Gao J, Shen H-W, Garcia A.[4] concerned about data decomposition problem: Their algorithm begins with rapidly skipping the empty cells and collecting the active cells by traversing the octree data structure. For the load balancing, a binary space partitioning scheme which recursively partitions the entire screen into tiles is used to ensure that each processor will generate approximately the same amount of triangles. Zhang H, Newman T[5] proposed an approach that focuses on trying to minimize the usage of the memory and optimizing disk I/O at the same time achieving a balanced load. First they partitioned the dataset into subfiles according to a non-overlapping interval of the range of potential isovalues, then the workload is estimated using a linear function to find the total number of cubes and the number of active cubes. They also used a hybrid granularity approach to optimize disk access. M. Tchiboukdjian and V. Danjean [6] proposed a cache-efficient parallel isosurface extraction algorithm. They take advantage of shared caches, and achieve the goal of deducing cache misses by having cores working on close data. To parallelize it, they divide the cells into some chunks and process them in parallel. To process one chunk, it also divides the cells into groups, one for each core.

In this paper, we show that generating a polygonal mesh of an isosurface for a large-scale dataset can be done in an efficient and scalable manner. The main purpose of our research is to generate a triangular mesh with high efficiency for the sequential algorithm and to get a linear speedup of the parallel algorithm. To achieve this goal, our algorithm works on two lookup tables[9], one for finding the vertices which exactly cut the isosurface, while the other is for looking up the vertex sequence for triangular facets to represent the isosurface. The crux of achieving good scalability is evenly distributing the

workload to each core to make sure that each thread works on approximately the same number of active cells in order to generate approximately the same amount of triangles. What is more, in order to achieve the maximum concurrency, our algorithm not only parallelize the generation of a triangular mesh for active cells, but also parallelizes the creation of background rectilinear grids for all cells, and each part gets its balanced workload.

The rest of the paper is organized as follows. In Section 2 we briefly describe the double table-based Marching Cubes algorithm. Section 3 presents the framework for our parallel algorithm. In section 4, we present the implementation issues and the experimental results. Finally we conclude with a summary of our results.

II. DOUBLE TABLE-BASED ALGORITHM

The sequential algorithm starts with creating the cubical cell of the sampled rectilinear grid one by one, and by examining the values in its corners, identifies the index of the two tables of the cell. Every cube has a same index of the two tables, range from 0 to 255. Then the algorithm uses the first predetermined table to get a list of vertices which cut the cube edges, and uses the same index to look up the second table to get the cut-vertex order of triangles to construct the local patches of the isosurface inside the current cubical cell. The union of the isosurface patches from all cubical cells makes up the final result. The exact position of the intersections of the isosurface with grid edges are determined by linear interpolation.

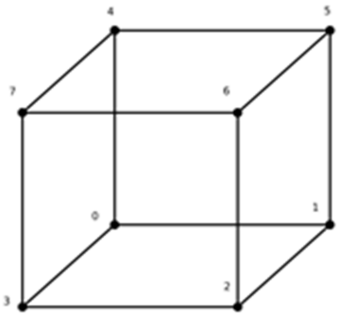
The main steps of the double table-based marching cubes algorithm are shown in the algorithm of Fig.1. The function TABLELOOKUP1(c) queries a manually constructed table with the key composed of eight bits, each bit corresponding to the result of the test, $F(x)>= \xi$ or $F(x)< \xi$, where ξ denotes the isovalue, in one of the eight corners of cube c. The function TABLELOOKUP2(c) queries a manually constructed table which is a two dimensional array, each row composed of 16 values, the number of values which is not -1 is a multiple of 3. Each 3 none -1 values are composed of one triangle's cut-vertex index. Each row represents the sequence of cut-vertices of triangles in one cube. The return value of this function is a set of triangles defined by their vertices.

Algorithm DoubleTableBasedMarchingCubes(I, ξ)
Input: I is a three-dimensional image, i.e., a three-dimensional scalar array of VTK file, and an isovalue $\xi \in \Re$
Output: A triangular surface M embedded in that interpolates the set $\{ x \in \Re^3 \mid F(x)=\xi\}$
1 $M \leftarrow \phi$
2 generate the background cubes G in $\Re^3$ along with a mapping $F: G \rightarrow \Re$ one by one
3 for each $c \in C$
4 for each point in G, compute table index by determining whether $F(x)>= \xi$ or $F(x)< \xi$
5 get a list L of cut-vertex by TABLELOOKUP1(c)
6 $M \leftarrow M \cup$ TABLELOOKUP2(c)
7 endfor
8 return M
Fig. 1. A high level description of the marching cubes algorithm based on two tables lookup.

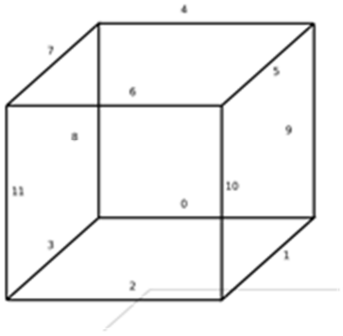
III. PARALLEL MARCHING CUBES ALGORITHM

A. Indexing Conventions

In our entire development the naming convention is suitable for any cube of the sampled grid. We call the points in the corners of *cube corners*, and the points in the intersection of the resulting isosurface with the edges of the generic cube *cut-vertices* (or simply *vertices*). The indexing convention for cube corners and cut-vertices is introduced in Fig. 2.



(a) indexing conventions for the *cube corners* of each cube.



(b) indexing conventions for *cut-vertices.*
Fig. 2

We will use indexing conventions of cube corners and cut-vertices to introduce indexing conventions of the two look up tables. The first table stores a one-dimensional array composed of 256 entries. The index can be calculated by eight bits representing eight corners' insideness. The insideness has two cases, if $F(x)>= \xi$ it means this corner is outside the isosurface; if $F(x)< \xi$, it means the corner is inside the isosurface. The eight insideness completely determine the triangulation inside the cube. Each entry is a 12-bits representation which cut-vertices cut the edge of cubes by isosurface. For example, if the value at corner 0 and corner 3 are below the isosurface value and values at all other corners are above the isovalue, the index is decimal representation of

00001001, which is 9. The entry of this index is hexadecimal representation of 100100000101, which is 0x905. From this entry we know that vertices 0, 2, 8, 11 cut the edges of the cube. The second table stores another two-dimensional array also composed of 256 entries. Each entry consists of 16 integers. If the 8 corners of a cube are all inside or outside the surface, the 16 integers are all -1. Otherwise, the number of non-negative integers is a multiple of 3, and each triple represents an output triangle. For the same example as above, the entry of index 9 is {0,11,2,8,11,0,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1}, therefore we can get *triangle {0 11 2}* and *triangle {8 11 0}*. Fig. 3 shows these two triangles.

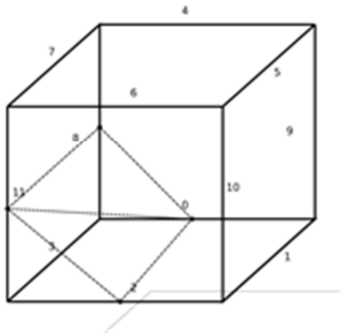


Fig. 3. Two triangles generated by two tables when corner 0 and corner 3 are inside the isosurface

B. Parallel Double Table-Based Algorithm

The fundamental problem of achieving the maximum concurrency of parallel double table-based marching cubes algorithm is parallelizing all the steps of it. And to achieve linear speedup, we must distribute workload evenly in every step of our algorithm. The first step of our algorithm is parallel generation of the rectilinear grids of the three-dimensional medical image. Instead of evenly dividing the dataset layers, we partition the total number of cubes into clusters in which the number of cubes are equal except the last one. But the difference of maximum number of cubes between the last processor and other ones is not more than the maximum number of threads. That means, even if we use 64 threads, the largest difference is 63. Compared to the large size of the data sets, for example, an image whose size is 1000*1000*1000, this can be deemed as a completely balanced workload. After generating the background grids we calculate the number of index for each cube which is the same for the two tables.

For the next step, we use the index calculated in the previous one to lookup two tables then to generate the output triangles. If the eight corners of a cube are all inside or outside the surface, there will be no output triangles. If the index of one cube is either 0 or 255, that means this cubes never generate output trangles, we call it non-active cube, otherwise, we call it active cells. The active cubes in one image is only a small part of the whole image. To gain the highest efficiency of our algorithm, we can skip these steps by always checking whether its index is 0 or 255. The problem is the threads which got the approximately same number of cubes can not get the same number of active cells. To re-distribute the active cubes evenly to all threads, we store the indices into a local vector for each thread, and then calculate the average number of active cubes for each thread. Similar to the first step, the difference between the maximum number of active cubes is not more than the maximum number of threads. Fig. 4 shows the main steps of the parallel double table-based marching cubes algorithm.

Algorithm ParallelDoubleTableBasedMarchingCubes(I, ξ)
Input: I is a three-dimensional image, i.e., a three-dimensional scalar array of VTK file, and an isovalue $\xi \in \Re$
Output: A triangular surface M embedded in that interpolates the set $\{ x \in \Re^3 \mid F(x)=\xi \}$
1 $M \leftarrow \phi$
2 each thread gets approximately the same number of cubes, in parallel generates the background cubes G in $\Re^3$ along with a mapping $F: G \rightarrow \Re$ one by one
3 for each $c \in C$
4 for each point in G, compute table index by determining whether $F(x) >= \xi$ or $F(x) < \xi$ and stores them in a local vector
5 calculate and distribute the average number of active cubes for each threads, each active cube gets a list L of cut-vertices by TABLELOOKUP1(c)
6 for each thread, $M' \leftarrow M' \cup$ TABLELOOKUP2(c)
7 endfor
8 M = union M'
9 return M

Fig. 4. A high level description of the parallel marching cubes algorithm based on two tables lookup

IV. Experimental Evaluation

A. Implementation

This parallel algorithm is written using the C++ programming language. It uses POSIX Threads Programming to implement parallelism on shared memory multiprocess architectures. We chose to use a vector instead of a three-dimensional array to store the volume data to gain a better load balancing. We try to avoid the use of any global array or vector to guarantee the efficiency of the implementation.

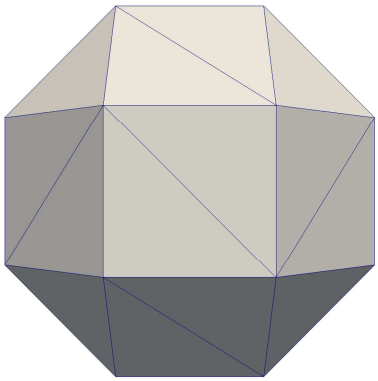
B. Experimental results

For the experimental evaluation of our code, we used the CRTC work station (CPU Intel(R) Xeon(R) X5690@ 3.47GHz Ubuntu 64-bit operation system). It has two sockets, and each socket has 6 cores and 48GB of memory. The experiments involving sequential software were performed on one of its cores.

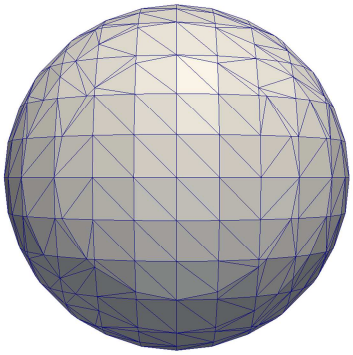
One control when polygonising a field where the values are known everywhere in the field is the resolution of the sampling grid. The isosurface can be generated as either course or fine approximation depending on the resolution required. Table 1 shows different data sizes and the corresponding numbers of triangles generated by our algorithm. Fig. 5. shows a sphere at different grid size.

TABLE I
DIFFERENT DATA SIZE AND THE CORRESPONDING NUMBER OF TRIANGLES GENERATED BY OUR ALGORITHM

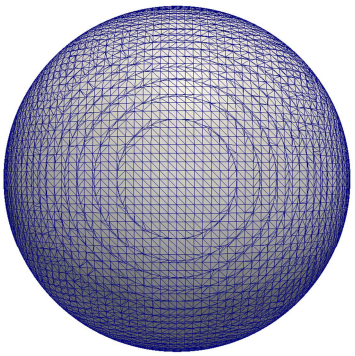
Data Size	Number of Triangles
3*3*3	44
10*10*10	824
50*50*50	23,288
500*500*500	2,355,560



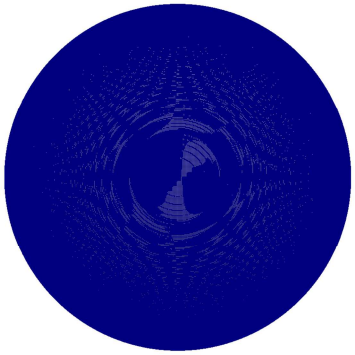
(a) The data size is 3*3*3.



(b) The data size is 10*10*10.



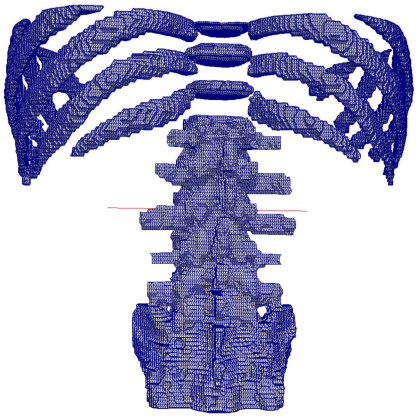
(c) The data size is 50*50*50.



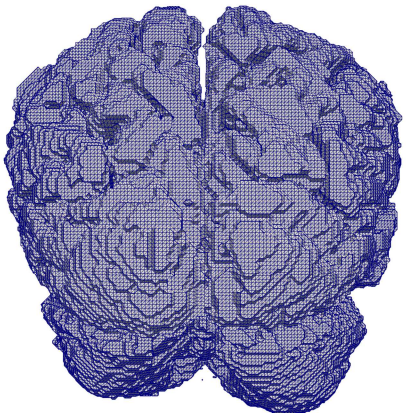
(d) The data size is 500*500*500.

Fig. 5. A sphere at different grid resolutions.

In Fig. 6 we show some isosurfaces extracted by the double table-based algorithm. These are relatively small images, their resolution is 256*256*129 and 256*256*159.



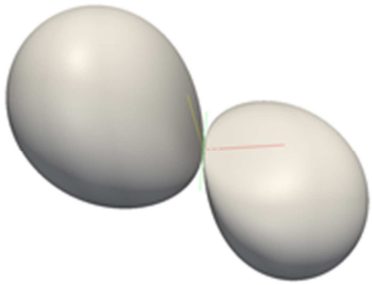
(a) The isosurface extracted by the double table-based algorithm on SPL Abdominal Atlas[7].



(b) The isosurface extracted by our double table-based algorithm on SPL-PNL brain atlas[8].
Fig. 6

We generate a very large data set using the sphere model to test the performance of our parallel algorithm. The max data size is 1800*1800*1800. We also generate our own model—double water drop model. The data size is 1000*1000*1000. The two models are showed in Fig. 7. Table2 shows the running time of our parallel algorithm when we use different number of threads of the two models. This time includes generating rectilinear grids, partitioning the image and extracting the triangles. Fig. 8 presents the scaled speedup evaluation using this model. From this figure we can see that our speedup with 12 threads and below is linear.

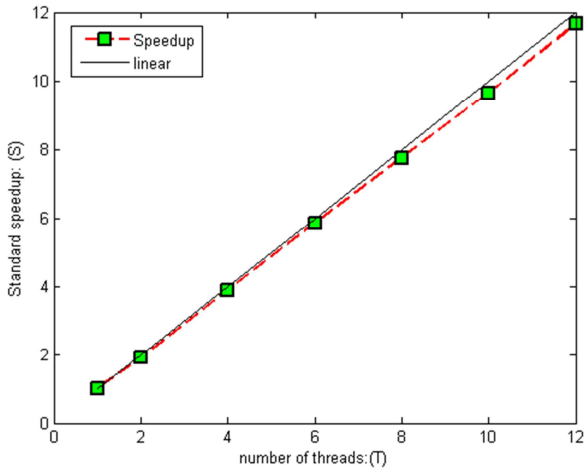
(a) sphere model.



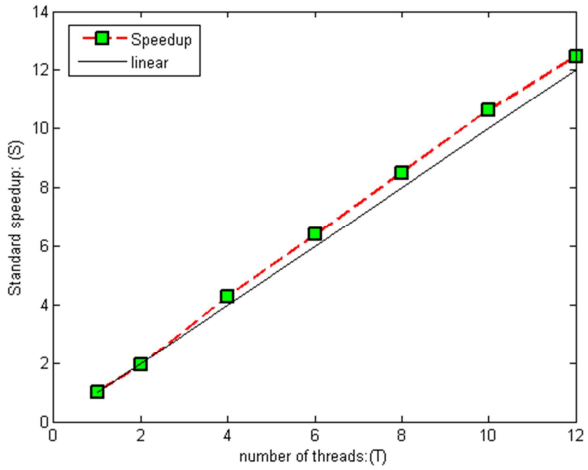
(b) double water drop model.
Fig. 7

TABLE II
RUNNING TIME OF OUR PARALLEL ALGORITHM WHEN WE USE DIFFERENT NUMBER OF THREADS OF SPHERE MODEL DOUBLE WATER DROP MODEL

	Sphere model	Double water drop model
1	45.6721	35.9133
2	23.0221	18.6597
4	11.2443	8.3968
6	7.4754	5.6198
8	5.6422	4.2202
10	4.5337	3.3819
12	3.7447	2.8782



(a) the scaled speedup measurements of sphere model.



(b) the scaled speedup measurements of double water drop model.
Fig. 8

V. CONCLUSION

We presented an algorithm and an implementation for the parallel double table-based marching cubes algorithm. Isosurface extraction is done in parallel by multiple threads on shared memory, and the workload can be partitioned evenly among the threads. Our experimental results show near linear speedup of the code and its ability to generate over a billion triangles only in a few seconds. Our algorithm can be easily combined with any polygon rendering algorithm.

REFERENCES

[1] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. SIGGRAPH Comput. Graph., 21(4):163-169, August 1987.

[2] Gerstner T, Rumpf M. Multiresolutional parallel isosurface extraction based on tetrahedral bisection. In: Chen M, Kaufman A, Yagel.R, editors. Volume graphics. London: Springer; 2000. p. 267–78.

[3] Chiang Y-J, Farias R, Silva C, Wei B. A unified infrastructure for parallel out-of-core isosurface extraction and volume rendering of unstructured grids. In: Proceedings of IEEE parallel and large data visualization and graphics, San Diego, 2001. p. 59 – 66.

[4] Gao J, Shen H-W. Parallel view dependent isosurface extraction using multi-pass occlusion culling. In: Proceedings of IEEE parallel and large-data visualization and graphics, San Diego, 2001.p. 67–74.

[5] Zhang H, Newman T. Efficient parallel out-of-core isosurface extraction. In: Proceedings of symposium on parallel and large-data visualization and graphics, Seattle, 2003. p. 9–16.

[6] M. Tchiboukdjian and V. Danjean. Cache-Efficient Parallel Isosurface Extraction for Shared Cache Multicores. Eurographics Symposium on Parallel Graphics and Visualization , 2010.

[7] I. Talos, M. Jakab, R. Kikinis, and M. Shenton. SPL-PNL brain atlas. http: //www.spl.harvard.edu/publications/item/view/1265, March 2008.

[8] Talos I-F., Jakab M., Kikinis R. SPL Abdominal Atlas. http://www.spl.harvard.edu/publications/item/view/1918

[9] Based on tables by Cory Gene Bloyd. http://local.wasp.uwa.edu.au/~pbourke/geometry/polygonise

Estimating Lower Bounds on the Length of Protein Polymer Chain Segments using Robot Motion Planning

Andrew McKnight, Jing He, Nikos Chrisochoides and Andrey Chernikov
Department of Computer Science
Old Dominion University
Norfolk, VA, USA
{*amcknigh, jhe, nikos, achernik*}*@cs.odu.edu*

***Abstract*—Finding the 3D structure a protein will assume given only its amino acid sequence remains largely an open question in bioinformatics. New developments have incorporated 3D images–"density maps"–of molecules from electron microscopy, but this presents its own problems. We seek to measure the lengths of segments of protein polymer chains associated with specific regions of the density map. We briefly discuss past efforts, and introduce an analog to robot motion planning as a novel approach. We will show this new approach's superiority through complexity analysis, and present some experimental results in the 2D case, leaving the 3D case to future work.**

***Keywords*-simplification; bounded; skeleton; shortest-path; graph**

I. Introduction

In the field of proteomics, an important question is how to efficiently determine the structure of a protein given its amino acid sequence: determining the phenotype of the protein from just the genotype, usually using machine learning methods. This approach has a vast solution space and exponential complexity. One promising advance in this area includes data from electron microscopy images of the protein molecules, the latest development being cryogenic electron microscopy (cryoEM). This approach can aid in discovering the structure and topology being sought and hopefully provide insight into the mechanisms that transform a given amino acid sequence into a functioning structure.

The microscopes produce three dimensional grayscale images whose voxels represent measured electron densities at the corresponding locations in space. These images are used in two ways to derive the protein's topology:

1) α-helix detection: Helixhunter [10] was previously used to determine the location of any α-helices, one of two secondary structure elements (SSEs) commonly found in proteins. More recently, we have developed and used our own gradient-based tool to detect the helices [19].
2) skeletonization: the Gorgon software package [16] produces a set of line segments and planes making up the basic "shape" described by the 3D density map in a process called morphological thinning [14,15]. Gorgon requires an input parameter; we have developed a tool that requires no input parameter, but that produces slightly more connected skeletal structures [18].

Kamal et al.'s topology ranking algorithm [3] attempts to find the correct pathway through the skeleton and α-helices that the amino acid sequence takes. They showed how to shrink the original solution space of $O(n!2^n)$ possible topologies, where n is the number of α-helices, to $O(n^2 2^n)$ using dynamic programming methods. The algorithm is able to narrow the possible topologies by comparing the distances between detected α-helices and the number of amino acids that are known to lie between them (it is generally accepted that the distance between two amino acids is between 3.5 and 3.8 Å).

Topology matching relies on knowing the distances between α-helix endpoints along the protein backbone, seen in Figure 1. We can use the skeleton to approximate the location of the polymer chain, and therefore measure its length, but it often contains many right angles, inflating the length estimates we derive from them. As a countermeasure, we employed the Douglas-Peucker line simplification algorithm [4]. While this is a widely accepted algorithm for these purposes, it requires an input threshold which affects its output, hence another degree of freedom in our own solution space. Other line simplification algorithms exist, claiming optimality under different metrics and independence from this input parameter [5,6,7,8].

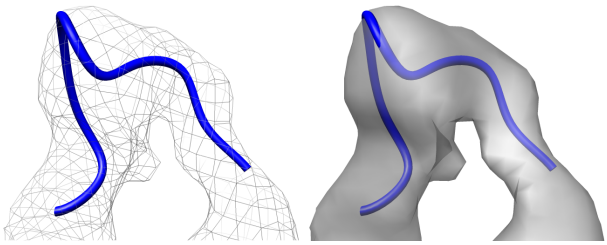


Figure 1. The tunnel-like partition of space induced from all matrix values over a threshold value in the 3D electron density map, and the associated segment of protein backbone whose length we want to know.

There are several limitations with our previous approach,

both theoretical and practical. First, by constraining our pathfinding to the skeleton of an image, we dramatically reduce the size of our search space. A protein turn may twist around in the general areas of high electron density as encoded in the image–the skeleton represents one of many possible such paths. By creating a roadmap containing all locations within an iso-surface in the image, we are able to try many different paths. We also avoid the need for input parameters for skeletonization and curve simplification, by working directly with the density map values, in exchange for one parameter to construct the iso-surfaces.

II. Approach

Our approach is analogous to the robot motion planning problem [2]. This can be done in any number of dimensions–and the protein problem is in three dimensions–but for purposes of simplicity we will present our approach in only two. Our robot is a point with no size, and its work space and configuration space (which are identical for point robots) are the area inside the iso-surface defined in the image, as illustrated in Figure 1 in the 3D case. Such a tunnel can be obtained by removing all voxels from the 3D image under a certain threshold value.

The problem, as stated formally, is the following: given a simple polygon $\mathcal{P}$ subdividing the plane and a start point s and end point f, find the shortest path from s to f that lies completely inside the polygon. $\mathcal{P}$ and the robot's configuration space $\mathcal{C}$ are both embedded in $\mathbb{Z}^2$, hence our solution will also contain only points in $\mathbb{Z}^2$. In future work, we may choose to subdivide the intervals between points to obtain a finer discretization of the tunnel interior for more precise shortest path estimates. Figure 2 illustrates a simple case.

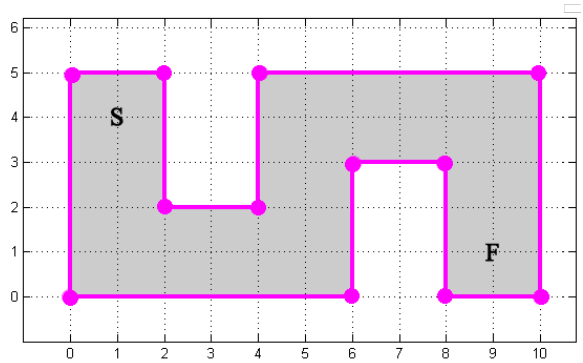


Figure 2. A 2D tunnel embedded in $\mathbb{Z}^2$ with its vertices and start and end configurations. The shaded region represents the configuration space we will work in.

Beginning with the grayscale density image $\mathcal{I}_G$, we generate a binary image $\mathcal{I}_B$ by setting all voxels with values lesser than some threshold $\mathcal{T}$ value to 0, and those voxels with values higher than the threshold to 1. The union of edges between pixels of opposite binary value forms an *iso-surface* polygon $\mathcal{P}$ in 2D, and the threshold value that produced $\mathcal{I}_B$ is called the *iso-surface threshold*. The pixels inside the iso-surface can be readily discerned from the binary image, by simply taking all entries with a value of 1.

Next, we construct our roadmap graph $\mathcal{G}$ by examining all pairs of interior pixels and testing whether a straight line segment between them intersects any edge in $\mathcal{P}$. If there is no intersection, then this is a possible path to take through $\mathcal{P}$, and two segments are added to $\mathcal{G}$ with the Euclidean distance between endpoints as the edge weight: one in each direction. When complete, $\mathcal{G}$ represents our entire search space for all possible locations within the iso-surface. Figures 3 and 4 summarize this process.

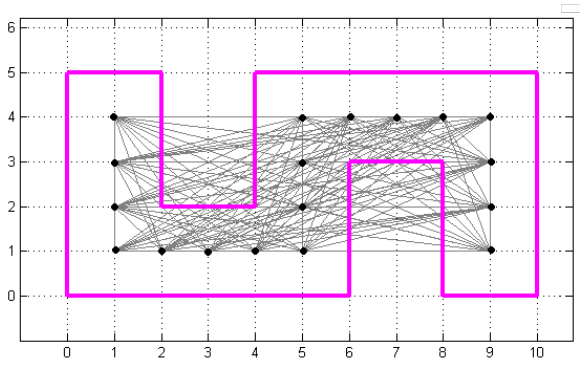


Figure 3. The initial tunnel voxel graph.

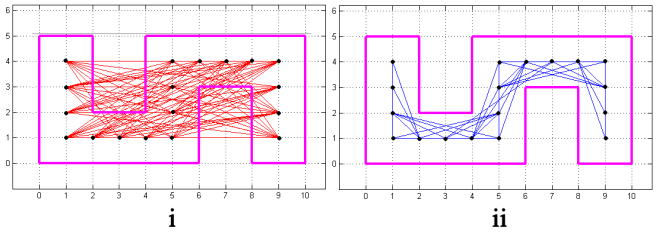


Figure 4. i) The removed edges that intersect the tunnel boundary. ii) The legal moves our robot can make to find the shortest path between s and f.

Using our search space, we would like to find the shortest possible path, representing the lower bound on the length of the protein's turn. Our roadmap is a positively weighted, directed graph, and so we use Dijsktra's algorithm to find all shortest paths from s, and reconstruct the path that ends at f. A simple case is shown in Figure 5. There are several situations where the choice of endpoints differ:

1) An endpoint lies inside $\mathcal{P}$. We simply use the endpoint.
2) An endpoint lies outside $\mathcal{P}$. We search through $\mathcal{P}$'s interior voxels for the one closest to the endpoint in terms of Euclidean distance.

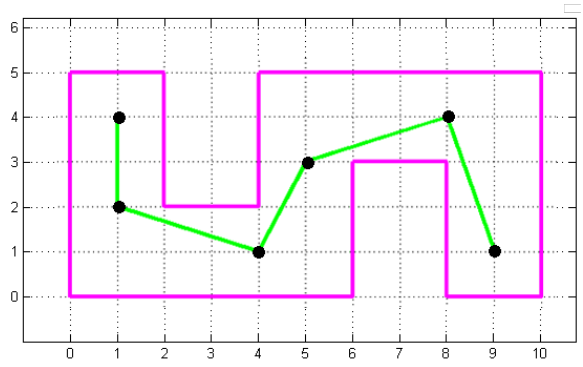


Figure 5. The shortest path between s and f that lies inside the tunnel.

The distances between vertices can be easily computed with the result and summed to obtain our lower bound estimation on the length of the protein polymer chain. Figure 6 summarizes the process:

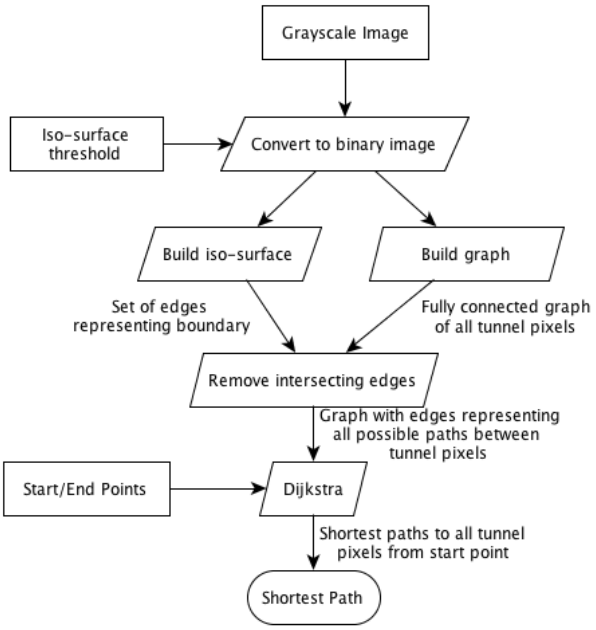


Figure 6. The complete algorithm.

III. Analysis

Our 2D images are square matrices of size $s \times s$, so there are exactly $n = s^2$ pixels. The binary images we produce are sparse binary matrices, with $\hat{n} = O(n)$ vertices, usually much less than n. Generating the $\mathcal{I}_B$ requires visiting each pixel once and so is $\Theta(n)$. Collecting $\mathcal{P}$'s interior pixel also takes linear time w.r.t. n. To collect $\mathcal{P}$'s edges from $\mathcal{I}_B$, each pixel is again visited once, and its four adjacent neighbors are inspected, again a linearly asymptotic operation with a constant number of operations at each pixel.

The most intensive operations are composing the roadmap by testing for intersections between graph edges and $\mathcal{P}$'s edges, and searching the roadmap for the optimal shortest path from s to f. By testing all possible pairs of interior pixels, we must make $O(n^2)$ comparisons. At each comparison, we are testing the intersection between the line segment l and any edge in $\mathcal{P}$, which can be checked in constant time. Dijkstra's algorithm is quadratic in the number of vertices in the graph, and so is also $O(n^2)$. Reconstructing the path afterwards is trivial in comparison.

The overall complexity of the algorithm is dominated by the $O(n^2)$ terms, so is quadratic w.r.t. the number of voxels in the cube. Figure 7 summarizes the complexities of the algorithm's components.

shortestPath($\mathcal{I}_G$, s, f)

1) convert $\mathcal{I}_G$ to $\mathcal{I}_B$ $\mathcal{I}_B$	$\Theta(n)$
2) construct $\mathcal{P}$'s edges	$\Theta(n)$
3) gather $\mathcal{G}$'s vertices	$\Theta(n)$
4) construct $\mathcal{G}$	$O(n^2)$
5) find intersections between $\mathcal{P}$ and $\mathcal{G}$	$O(n^2)$
6) Dijkstra's algorithm	$O(n^2)$

Figure 7. The algorithm with time complexities.

IV. Results

Testing was performed with a 1.7 GHz quad-core Intel i5 and 4 GB 1333 MHz DDR3 RAM, with cases ranging in size from $9 \leq a \leq 248$. A quadratic trend is evident in the observed test cases, as can be seen in Table 1 and Figure 8. Table 1 shows the amount of tunnel voxels per case, runtimes, and multiplication factors between subsequent cases' a and runtime values.

Table I
Runtimes with respect to a.

Case (i)	a	Runtime[1]	$\frac{a_i}{a_{i-1}}$	$\frac{\text{Runtime}_i}{\text{Runtime}_{i-1}}$
0	9	0.48	-	-
1	10	0.86	1.11	1.79
2	10	0.6	1.0	0.7
3	11	0.69	1.1	1.15
4	12	0.84	1.09	1.22
5	15	0.186	1.25	0.22
6	16	0.226	1.07	1.21
7	18	0.179	1.13	0.79
8	19	0.208	1.06	1.16
9	22	1.64875	1.16	7.93
10	25	1.64755	1.14	1.0
11	26	0.623	1.04	0.38
12	27	0.399	1.04	0.64
13	48	6.65008	1.78	16.67
14	94	24.6529	1.96	3.71
15	248	262.386	2.64	10.64

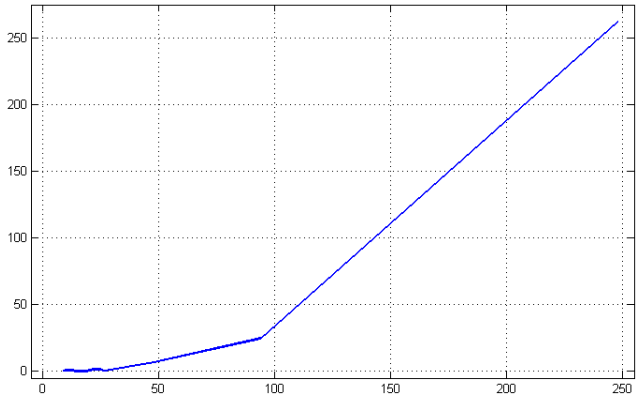


Figure 8. Runtimes with respect to a.

We implemented the algorithm in C++, utilizing the CGAL [15] library for applicable geometric computations and the Boost library [14] for graph operations. Some basic geometries were supplied as test cases, as well as some larger ones representative of the molecular structures that the algorithm is meant to evaluate, all of which can be seen in Figures 9-16. Cases 0-4 show the effect of eroding the corner of an l-shaped polygon; also tested were polygons with xy-, x-, y- and no monotonicity. Square-like and start-like polygons are both present. All the test cases are in two dimensions. However, our actual interest is in 3D molecules, and therefore 3D tunnels and corresponding paths. We are nearing completion of expanding our algorithm to 3D for work on actual density maps.

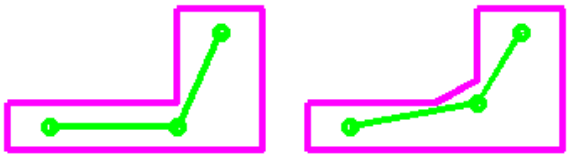


Figure 9. Left: case i = 0. Right: case i = 1.

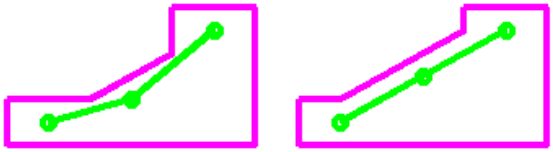


Figure 10. Left: case i = 2. Right: case i = 3.

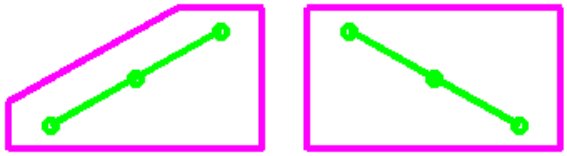


Figure 11. Left: case i = 4. Right: case i = 5.

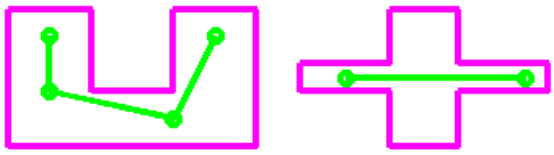


Figure 12. Left: case i = 6. Right: case i = 7.

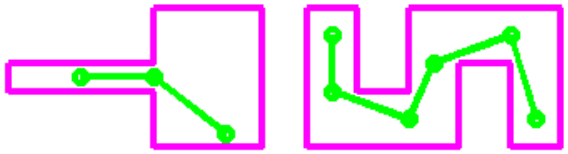


Figure 13. Left: case i = 8. Right: case i = 9.

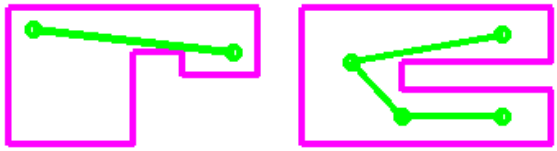


Figure 14. Left: case i = 10. Right: case i = 11.

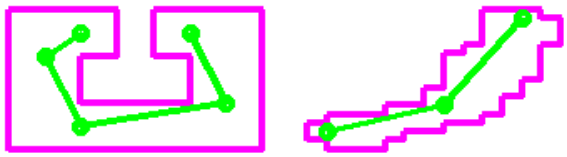


Figure 15. Left: case i = 12. Right: case i = 13.

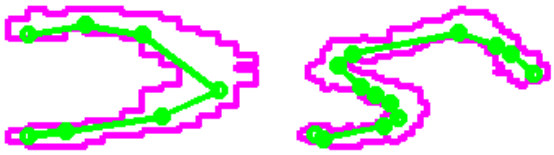


Figure 16. Left: case i = 14. Right: case i = 15.

We also included, as a real-world example, a two-dimensional slice from a density map. The map was generated using EMAN [20], by extracting the turn residues from a PDB atomic structure file and supplying them as input, to produce a synthetic density map representing only the turn. The endpoints of the path are the corresponding endpoints from the detected helix curves, as detected from a density map generated from the surrounding helix-turn-helix motif the turn in which the actual turn is located. We are currently using these synthetic maps to reduce the noise present, which is usually seen in real data, for initial development of this new technique. Figure 17 shows the results from this case.

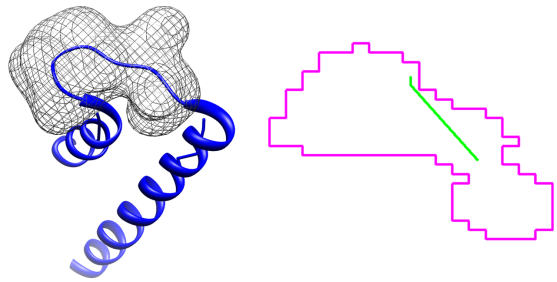


Figure 17. Left: a helix-turn-helix motif from PDB 1R1H in blue, with the density map generated using the turn residues as the grey mesh. Right: The iso-surface boundary for the median density value in magenta, with the shortest path between the two helix endpoints in green.

V. Conclusion

Robot-motion planning principles show promise in estimating the minimum length of a protein chain between two helices given only the 3D density map for the molecule. It replaces several steps: skeletonization, all-pairs shortest path computation and line simplification, thereby reducing the asymptotic complexity of the general problem's solution. It is readily translatable into the 3D case, which we have left for further study. Overall, it is a more elegant solution to the problem of estimating the lower bounds of polymer lengths from 3D images.

Acknowledgment

The authors would like to thank Dong Si, Lin Chen and Kamal al-Nasr of ODU for their work in related areas of the protein topology matching problem, and for the software they have provided. They would also like to acknowledge the authors of the CGAL and Boost libraries.

References

[1] Thomas H. Cormen et al, *Introduction to Algorithms*, 3rd ed. Cambridge, Massachusetts: The MIT Press, 2009.

[2] Mark de Berg et al, *Computational Geometry: Algorithms and Applications*, 3rd ed. Berlin: Springer-Verlag, 2008.

[3] Kamal al Nasr et al, *Ranking Valid Topologies of the Secondary Structure Elements Using a Constraint Graph*, Journal of Bioinformatics and Computational Biology 9(3), 2011, pp. 415-430.

[4] John Hershberher and Jack Snoeyink, *Speeding Up the Douglas-Peucker Line-Simplification Algorithm*, Proc. 5th Intl. Symp. on Spatial Data Handling, 1992, pp. 134-143.

[5] Pankaj K. Agarwal et al, *Near-Linear Time Approximation Algorithms for Curve Simplification*, Algorithmica 43, 2005, pp. 203-219.

[6] Prosenjit Bose et al, *Area-Preserving Approximations of Polygonal Paths*, Journal of Discrete Algorithms 4, 2006, pp. 554-566.

[7] Wang Xiao-li and Zhang De, *Selecting Optimal Threshold Value of Douglas-Peucker Algorithm Based on Curve Fit*, First Intl. Conf. on Networking and Dist. Computing, 2010.

[8] Veregin, Howard, *Line Simplification, Geometric Distortion, and Positional Error*, Cartographica 36(1), 1999, pp. 25-39.

[9] Bernard Chazelle et al, *Algorithms for Bichromatic Line-Segment Problems and Polyhedral Terrains*, Algorithmica 11, 1994, pp. 116-132.

[10] A. Dal Palu et al, *Identification of α-Helices from Low Resolution Density Maps*, Comp. Syst. Bioinformatics Conf., 2006, pp. 89-98.

[11] Khalid Saeed et al, *K3M: A Univeral Algorithm for Image Skeletonization and a Review of Thinning Techniques*, Int. J. Appl. Math. Comp. Sci. 20(2), 2010, pp. 317-335.

[12] Sasakthi S. Abeysinge et al, *Segmentation-free Skeletonization of Grayscale Volumes for Shape Understanding*, SMI 2008, pp. 63-71.

[13] Ju, Tao, Matthew L. Baker and Wah Chiu, *Computing a Family of Skeletons of Volumetric Models for Shape Description*, Computer Aided Design 39(5), 2007, pp. 352-360.

[14] Siek, Jeremy G., Lie-Quan Lee and Andrew Lumsdaine, *The Boost Graph Library: User Guide and Reference Manual*, Addison-Wesley Professional, 2001.

[15] NA. *CGAL: User and Reference Manual*, Release 4.1, October 2012. Retrieved from http://www.cgal.org/Manual/latest/doc_pdf/cgal_manual.pdf

[16] NA. *The Gorgon Project*, Retrieved from http://gorgon.wustl.edu/index.php

[17] Andrew McKnight et al, *CryoEM Skeleton Length Estimation using a Decimated Curve*, IEEE BIBM CSBW, 2012, pp. 109-113.

[18] Nasr KA, Chen L, Si D, Ranjan D, Zubair M, He J, *Building the Initial Chain of the Proteins through De Novo Modeling of the Cryo-Electron Microscopy Volume Data at the Medium Resolutions,* ACM Conference on Bioinformatics, Computational Biology and Biomedicine, Orlando, FL 2012.

[19] Si D, Ji S, Nasr K, He J, *A machine learning approach for the identifcation of protein secondary structure elements from electron cryo-microscopy density maps.*hskip 1em plus 0.5em minus 0.4emBiopolymers 2012, 97(9):698-708.

[20] Ludtke S, Baldwin P, Chiu W, *EMAN: semiautomated software for high-resolution single-particle reconstructions*, J Struct Biol 1999, 128:82-97.

Simulation of the Localized 3-dimensional Reconstruction for Electron Cryo-tomography

Dong Si[1], Hani Elsayed-Ali[2], Wei Cao[2], Olga Pakhomov[3], Howard White[4], Jing He[1]

***Abstract* — Recent advances in electron cryo-tomography (cryoET) allows thin samples such as macromolecular complexes and small bacterial cells to be imaged. In order to reconstruct a three-dimensional image, computation is needed to combine 2-dimensional images taken from a number of different viewing orientations. In this paper, we focus on a special situation in which the biological interest is in a local region of an object that has a protrusion. We propose an idea to cut computationally in the region of interest at the protrusion. We propose a localized crop-cone model whose location is predictable in each 2D projection image once the viewing orientation of the projection is known. A weighting procedure was implemented in the 3D reconstruction based on the protruding feature and the projection orientation. Our simulation demonstrates that the localized cropping and reconstruction resolves the local features better than the global reconstruction.**

***Index Terms* — cell, electron cryo-tomography, 3D image reconstruction, simulation, modeling.**

I. Introduction

Electron cryo-tomography (cryoET) is becoming a major biophysical technique to study the three-dimensional (3D) structures of large cellular components [1, 2]. The Electron Microscopy Data Bank (EMDB) [3] currently archives 1794 (as of March 20, 2013) three-dimensional images with resolution from 3.1Å to 97Å resolution. Electron cryo-microscopy (cryoEM) uses similar principle as that of cryoET, but it generally works with smaller and more homogeneous biological samples. The targeted molecular complexes in cryoEM are purified from the cells while the targets of cryoET are generally inside the cell during the imaging to represent the in-vivo snapshot. CryoEM technique has been successfully used to study complexes such as viruses [4, 5], ribosome [6, 7], GroEL [8], and membrane bound calcium channels [9]. The success of cryoEM has demonstrated the theoretical capability of cryoET, but it also illustrated the practical difficulty of dealing with the large sizes of the cellular components in cryoET.

Correspondence to Jing He: jhe@cs.odu.edu

[1]Department of Computer Science, Old Dominion University
[2]Applied Research Center, Department of Electrical and Computer Engineering, Old Dominion University
[3]Frank Reidy Research Center for Bioelectrics, College of Health Sciences, Old Dominion University
[4]Department of Physiological Sciences, Eastern Virginia Medical School

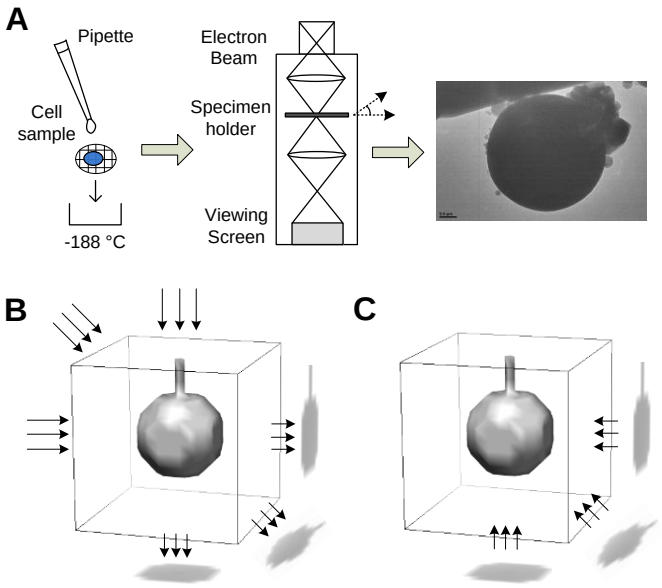


Figure 1. The principle of electron cryo-tomography and 3D reconstruction. (A) Freezing the specimen and to obtain 2D projections images using a cryoTEM; (B) An object is sampled by projection from various viewing orientations; (C) Reconstruction by back-projecting the projections along the original viewing orientations.

In the experiment of cryoET, biological specimen is quickly frozen using liquid nitrogen (Figure1 A). A transmission electron microscope (TEM) is used to collect the 2-dimensional (2D) projection images of the molecules from different tilted orientations, and the computational steps are involved to merge these 2D images into a 3D image of the molecules (Figure1 B and C).

Inspired by the success of cryoEM, cryoET has been identified as the most potential method in studying the sub-cellular arrangement of organelles. However, the current cryoET technology cannot produce high resolution 3D image of a mammalian cell due to its large size (> 1μm). The virus that has been resolved to 3.88Å by cryoEM is much smaller (~70nm) [4]). In general, the number of the 2D images that are needed to achieve the same resolution increases significantly as the size of the object increases. Current technology involves cutting the entire cell into thin sections using an ultra-microtome. CryoET is then applied to each section. Although new technology is being developed to cut the cells into thin sections, it is a technically challenging step. The cutting needs to be performed in liquid-nitrogen environment, and the static charge makes it difficult to manipulate the ultra-thin sections.

In this paper, we focus on a special situation in which the biological interest is in a local region of an object that has a protrusion. Instead of using a specialized knife to cut the object into thin slices, we propose an idea to cut computationally in the region of the protrusion that is of interest. We propose a localized crop-cone whose location is predictable in each 2D projection image once the viewing orientation of the projection is known. Our simulation demonstrates that the localized cropping and reconstruction resolves the local features better than the global reconstruction.

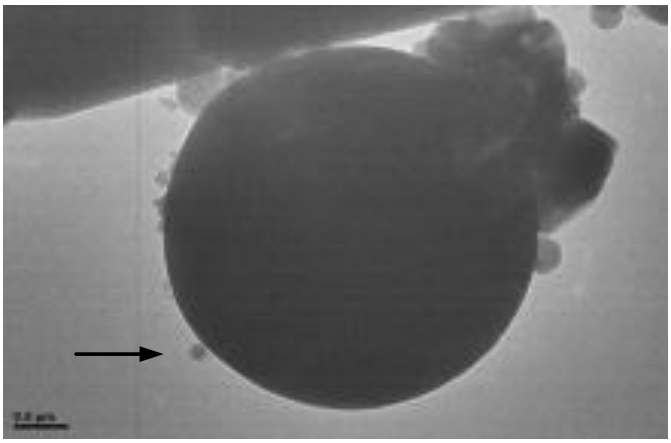


Figure 2. An image of a Jerkat cell taken by the cryo-Transmission Electron Microscope at the Applied Research Center of Old Dominion University.

II. METHODOLOGY

A. *Imaging of biological specimen using cryo-TEM*

The experiment was performed at the Applied Research Center of Old Dominion University using a cryo-equipped Transmission Electron Microscope. Solution containing Jerkat cells were applied to c-flat grids, and were plunged in liquid ethane that is in a bath of liquid nitrogen. The samples were transferred to a cryo-holder and were imaged by a JEM-2100F microscope (Figure 2).

B. *Overall of the localized reconstruction simulation*

The main difference between the standard cryoET and the localized cryoET lies in the 3D reconstruction step. Each 2D image generated by a TEM is a projection image of the object along the viewing orientation, since the image is formed by the electrons that penetrate the specimen object. The reconstruction step merges multiple 2D projection images into a 3D image according to the orientation of each projection. The principle of the 3-dimensional reconstruction is based on the principle of Radon theorem. It says that the projection image obtained by the microscope is the integral of the object along the electron beam direction (Figure3 A). Although the reconstruction can be performed in either real space (x, y, z) or in Fourier space, it is often performed in Fourier space for simplicity [10]. For the purpose of the localized reconstruction, we will consider real space reconstruction method, since it is easier to represent the location information in real space.

The central question to be answered in the localized reconstruction is how to distribute the projection information that comes from the entire object to different regions along the beam. The standard reconstruction method equally distributes the projection information along the direction of the beam. However, we are only interested in the projection information that comes from the local region of interest. We have developed a method to distinguish the projection information that belongs to the local region, and explored the use of weights to improve the localized reconstruction.

We simulated a 3D object which includes a protruding stick (Figure3 A). The main object is a ball with radius 4Å and centered at the image center, the protruding stick is a cylinder with radius 1Å and height 3Å that points upwards (Z direction). The entire object was saved as density MRC file with image size 15*15*15 and spacing 1 Å/pixel, each voxel in this image has density value equals to 1 if it's on the object or 0 if it's on the background. MRC is a file format for electron density that has become industry standard, it is a three-dimensional grid of voxels each with a value corresponding to the density of electrons [11]. This type of volumetric data can be viewed by almost every molecular graphic software, such as Chimera that was used in this paper [12].

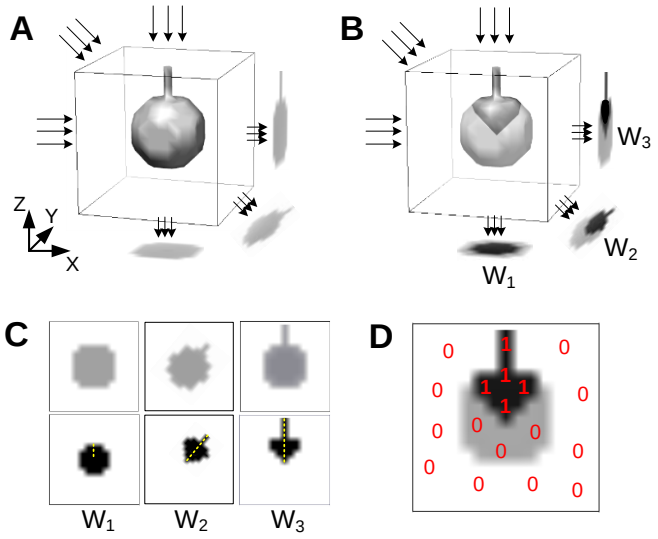


Figure 3. Localized reconstruction. (A) Standard reconstruction; (B) localized reconstruction; (C) Projection from different angles for the entire object (upper) and local region (lower), yellow dotted lines represent the weight for different projection orientation; (D) Local projection mask.

The core procedure of our localized 3D reconstruction can be divided into three steps: Local region generation, local projection mask and weighted 3D reconstruction.

C. *Local region generation*

Supposed we know the exact location of the local region that we are interested in. In order to reconstruct this local region instead of rebuilding the entire object, we modeled the local region as a cone, which includes the whole protruding part (Figure3 B). The apex of the local cone model is the same

as the center of the main object, although it can be another point in theory. The axis of the cone model passes through the protruding stick. The angle between the axis and the side of the cone is 45 degrees (Figure3 B). The purpose of modeling the cone region is trying to mask the local information out from the projection of entire object in the following step.

D. Local projection mask

The basic idea of 3D reconstruction is to build a 3D model through a series of 2D projection images along the projection direction. Our simulation of each 2D projection includes two steps. Firstly, we rotated the object into a number of sampled orientations using "e2proc3d" tool in EMAN package. EMAN is a suite of scientific image processing tools aimed primarily at single particle analysis [13, 14]. For each orientation, we integrated the density of this rotated 3D object along global Z direction to simulate the 2D projection image (Figure3 C upper row). Essentially, we tried to simulate the projection image that obtained by the microscope which is the integral of the object along the electron beam direction in real word. Since we have modeled the local region as a cone before, for sure we know where exactly the projection of that model is located on each projection of the entire object (Figure3 C lower row). For each orientation, the projections of the entire object and of the cone model were used to find the intersection. The points outside of the intersection were unrelated to the local region and were masked out (Figure3 D). The masking step helps us reconstruct only the local region instead of the entire object.

E. Weighted 3D reconstruction

Since we are more interested in the local protrusion on the object, the reconstruction process should emphasize more on the projections that have more information for the local protrusion. Intuitively, projections that were taken from the side directions should have more local information for the protrusion, while projections that were taken from the top or bottom directions should have almost no local information for the protrusion (Figure3 B and C). Based on this observation, we weighted each single projection by measuring the maximum distance from the apex of the cone model projection to any pixel on that projection (Figure3 C). So the weights for the different projection orientations in Figure 3 are $W_1 < W_2 < W_3$.

In the reconstruction step, each density on the 2D projection was equally distributed along the global Z direction into 3D space, and then we rotated this density distribution back to the original orientation. Essentially, we tried to simulate the reconstruction process which equally distributes the projection information along the direction of the beam (Figure1 C). Finally, all masked local projections were back-projected with their own weights, and then merged together into the object space:

$$3D\ localized\ reconstruction = \sum_{i=0}^{N} \Omega_i * \ W_i$$

Where N is the total number of masked local projections. Ω_i is the back-projecting density distribution for a particular rotation angle. W_i is the weight for that particular orientation.

III. RESULTS

We recently established the cryo-facility for an existing TEM at the Applied Research Center of Old Dominion University. The image in Figure 2 is one of our first set of images taken using this cryo-equipped TEM. The Jurkat cell (Figure 2) is about 3-4μm, a quite large specimen for the electron to penetrate. The small pieces appear to be attached to the surface of the cell is visible, although it is not clear what they might be.

We tested our localized reconstruction method using the simulated data described in the previous section. We sampled the projection for every 15 degree, from 0 to 360 degree rotationally in three dimensions. We simulated both the un-weighted and weighted reconstruction, each includes the global reconstruction that without the local projection masking and also the localized reconstruction. The modeling and simulation result in Figure 4 shows that our theoretical method of localized reconstruction is helpful for the local region reconstruction. Also, by introducing the weight to the method, more detail of the local region can be restored.

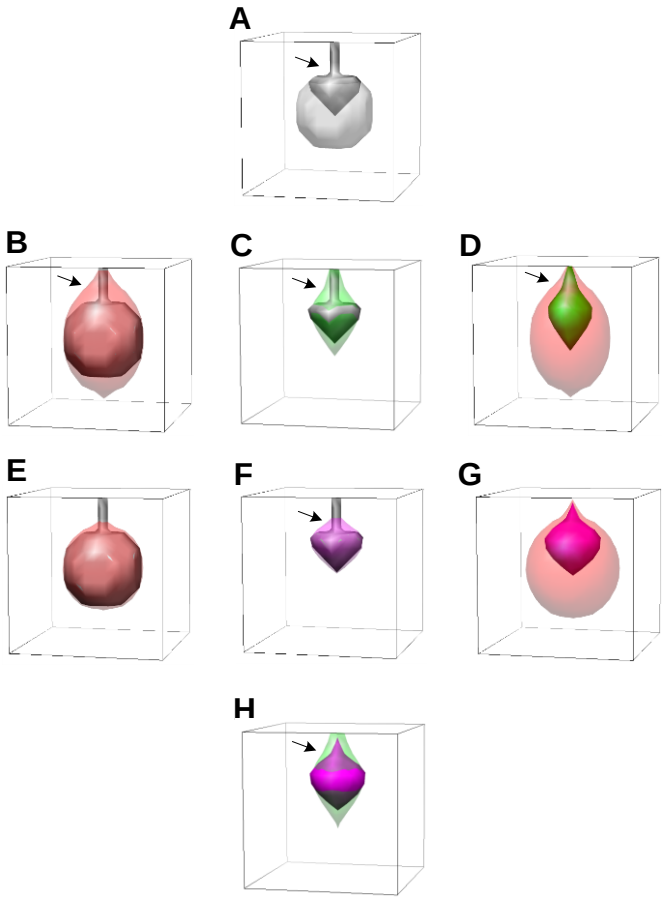


Figure 4. Localized reconstruction result. (A) Original object with the local cone model; (B) Global reconstruction (red) and the original object (gray); (C) Localized reconstruction (green) and the local cone model; (D) Global reconstruction (red) compared with localized reconstruction (green); (E) Weighted global reconstruction (red) and the original object (gray); (F)

Weighted local reconstruction (purple) and the local cone model (gray); (G) Weighted local reconstruction (purple) compared with weighted global reconstruction (red); (H) Weighted local reconstruction (purple) compared with un-weighted local reconstruction (green).

The compared result suggests that local features can be better restored by using localized reconstruction (Figure4 C and D, green) instead of using global reconstruction (Figure4 B and D, red). The localized reconstruction (Figure4 C and D, green) is more similar to the original local feature (Figure4 A) at the bottom of the protrusion. In order to compare fairly, we set the tip of the protrusion in the reconstructions to the ceiling of the reference box. We then evaluate and see which reconstruction resolves the bottom of the protrusion better. In addition, by introducing weight to the localized reconstruction, more detail of the protruding part can be reconstructed (Figure4 F, purple), which is better than the un-weighted result (Figure4 C, green). This is also shown by overlapping the two local reconstructions in Figure4 H.

Since the un-weighted method consider the projection from top and bottom directions the same as the projection from sides, the density distribution along the Z axis will be much stronger after back-projecting. That's why the reconstructed object for un-weighted method (Figure4 B-D) is stretched and much longer than the reconstructed object for weighted method (Figure4 E-G) in Z direction.

IV. CONCLUSION

Our preliminary results in this paper show that the localized 3D reconstruction can resolve more local features than the global reconstruction when the biological interest is in a local region of an object that has a protrusion. It uses a crop-cone to cut the area of interest from each 2D image in order to reduce the size of the object in the reconstruction. We showed that our orientation-based weighting method can be used to help improve the visualization of the local features. More research is needed to enhance the visualization of the local features and to tackle the challenges of large sizes of the cellular components.

ACKNOWLEDGMENT

We thank the help from Vitold E. Galkin for his guidance and testing of the cryoTEM. Funding for the research was provided by the Multidisciplinary Seed Fund of the Old Dominion University.

REFERENCES

[1] M. Faini, S. Prinz, R. Beck, M. Schorb, J. D. Riches, K. Bacia, B. Brugger, F. T. Wieland, and J. A. Briggs, "The structures of COPI-coated vesicles reveal alternate coatomer conformations and interactions," *Science*, vol. 336, pp. 1451-4, Jun 15 2012.

[2] W. Kukulski, M. Schorb, S. Welsch, A. Picco, M. Kaksonen, and J. A. Briggs, "Correlated fluorescence and 3D electron microscopy with high sensitivity and spatial precision," *J Cell Biol*, vol. 192, pp. 111-9, Jan 10 2011.

[3] C. L. Lawson, M. L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, G. van Ginkel, B. Devkota, I. Lagerstedt, S. J. Ludtke, R. H. Newman, T. J. Oldfield, I. Rees, G. Sahni, R. Sala, S. Velankar, J. Warren, J. D. Westbrook, K. Henrick, G. J. Kleywegt, H. M. Berman, and W. Chiu, "EMDataBank.org: unified data resource for CryoEM," *Nucleic Acids Res*, vol. 39, pp. D456-64, Jan.

[4] X. K. Yu, L. Jin, and Z. H. Zhou, "3.88 angstrom structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy," *Nature*, vol. 453, pp. 415-U73, May 15 2008.

[5] Z. H. Zhou, "Towards atomic resolution structural determination by single-particle cryo-electron microscopy," *Current Opinion in Structural Biology*, vol. 18, pp. 218-228, Apr 2008.

[6] M. Valle, R. Gillet, S. Kaur, A. Henne, V. Ramakrishnan, and J. Frank, "Visualizing tmRNA entry into a stalled ribosome," *Science*, vol. 300, pp. 127-30, Apr 4 2003.

[7] R. K. Agrawal, P. Penczek, R. A. Grassucci, Y. Li, A. Leith, K. H. Nierhaus, and J. Frank, "Direct visualization of A-, P-, and E-site transfer RNAs in the Escherichia coli ribosome," *Science*, vol. 271, pp. 1000-2, 1996.

[8] S. J. Ludtke, M. L. Baker, D. H. Chen, J. L. Song, D. T. Chuang, and W. Chiu, "De novo backbone trace of GroEL from single particle electron cryomicroscopy," *Structure*, vol. 16, pp. 441-8, Mar 2008.

[9] Serysheva, II, S. J. Ludtke, M. L. Baker, Y. Cong, M. Topf, D. Eramian, A. Sali, S. L. Hamilton, and W. Chiu, "Subnanometer-resolution electron cryomicroscopy-based domain models for the cytoplasmic region of skeletal muscle RyR channel," *Proc Natl Acad Sci U S A*, vol. 105, pp. 9610-5, Jul 15 2008.

[10] A. C. Kak, M. Slaney, and IEEE Engineering in Medicine and Biology Society., *Principles of computerized tomographic imaging*. New York: IEEE Press, 1988.

[11] R. A. Crowther, R. Henderson, and J. M. Smith, "MRC image processing programs," *J Struct Biol*, vol. 116, pp. 9-16, 1996.

[12] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, pp. 1605-1612, 2004.

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

[14] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "EMAN2: An extensible image processing suite for electron microscopy," *Journal of Structural Biology*, vol. 157, pp. 38-46, Jan 2007.

Using Constraints in Modeling the Protein β-Sheet Topology

Lin Chen[#], Kamal Al Nasr[$], Jing He[#*]

l

***Abstract*—Cryoelectron microscopy (CryoEM) is a technique to derive three dimensional structures for large molecular complexes. Although the individual amino acids are not visible at the mediate resolution, secondary structure elements (SSE) that are composed of groups of amino acids forming α-helices and β-sheets are visible. Combined with the secondary structure information from the protein sequence, the backbone of the protein can be modeled. A critical step in deriving the protein backbone is to determine the topologies that refer to the order and the direction of the secondary structures with respect to the protein sequence in a large searching space. In this study, we investigated the use of constraints that arise from the intrinsic propensities of β-sheet. A test using nine protein density maps shows that the constraints on β-strands improved the ranking of the true topology for all the nine cases.**

***Index Terms*—CryoEM, protein topology, SSE, top K, β-sheet, constraints**

I. Introduction

CRYOELECTRON microscopy (CryoEM) is a technique to determine the molecular structure for large protein complexes[1,2] . Traditional techniques for protein structure determination are X-ray crystallography and Nuclear Magnetic Resonance (NMR). Although both of them can provide the atomic level protein models, their intrinsic requirements limit their application[3].

In CryoEM experiment, solutions with the protein are quickly frozen at -188 ℃, Many two-dimensional (2D) images are collected using a cryo-equipped Transmission Electron Microscope (TEM). The 2D images taken from different viewing orientations are merged into a three-dimensional (3D) image, also referred as the molecular density map.

It is possible to detect secondary structure elements (SSEs), such as helices and β-sheets when the resolution of the three-dimensional image of the molecules is at the medium resolution such as 6-9 Å with algorithms such as HelixHunter[11], Sheetminer[12] and SSETracer[13]. It is still a challenging problem to derive the atomic structure from the 3D image of a protein [4-7]. The general procedure of building protein structure models from CryoEM is shown in Fig. 1. Density map of the specimen from CryoEM is segmented into domains (gray). Major SSEs, α-helices sticks (red rods) and β-sheet (green planes), can be located on this intermediate-resolution (6-9 Å) density map. Combined with the SSEs predicted from 1D sequence[8-11] (red bar for α-helices and green bar for β-strands), protein structure models can be generated by using reasonable topologies.

A core step to implement the above strategy is the determination of the native topology that assigns SSEs in sequence to the corresponding SSE sticks in density map correctly. Graphs (Fig. 2A) are employed to enumerate all possible topologies, in which each node represents an assignment of one SSE in the sequence to one secondary structure stick in the density map and the edge between nodes represents the possibility to have a loop with the suitable length. Ma[7] screened the geometrical unpreferred topologies with the loop constraint that erases the connections that loop length between nodes is less than the distance between corresponding stick side points, occupancy constraint that length of SSEs in sequence can not be significant different in length with its sticks in one node, geometrical filter that evaluates the scoring function value by doing statistics of secondary structure elements relative positions for PDB files in PDB bank[14]. This method enumerates all topologies and removes part of them and maximum stick number that can be evaluated is 8 which limits the application because the native topology is the last candidate to be evaluated in the worst case. Meiler[4-5] used the Monte Carlo method to enumerate the possible topologies and screen the unpreferred topologies with loop score, occupancy score and connectivity score, in which loop score and occupancy score are similar with Ma's constraints. Connectivity score detects the density between two sticks. The stronger the density is, the higher the possibility that connectivity exists between two nodes. This method increases the size of protein that can be predicted up to 10 SSEs by abandoning the detection for all possible topologies and cannot guarantee the native topology is contained in the scanned topology space. He[17] developed an algorithm to get the top K ranked topologies from the topology graph using Meiler's loop score, occupancy score and connectivity score. This method avoids enumerating all topologies and increases the size of protein accommodated in graph up to 30. Although this algorithm was designed for proteins including both α-helices and β-sheets, the method only has good performance on pure α-helices proteins due to the complexity of the structure of β-sheets.

[#]: Department of Computer Science, Old Dominion University, Norfolk, VA 23508
[$]: Department of Computer Science, Howard University
[*]: Correspondence to Jing He, e-mail: jhe@cs.odu.edu.

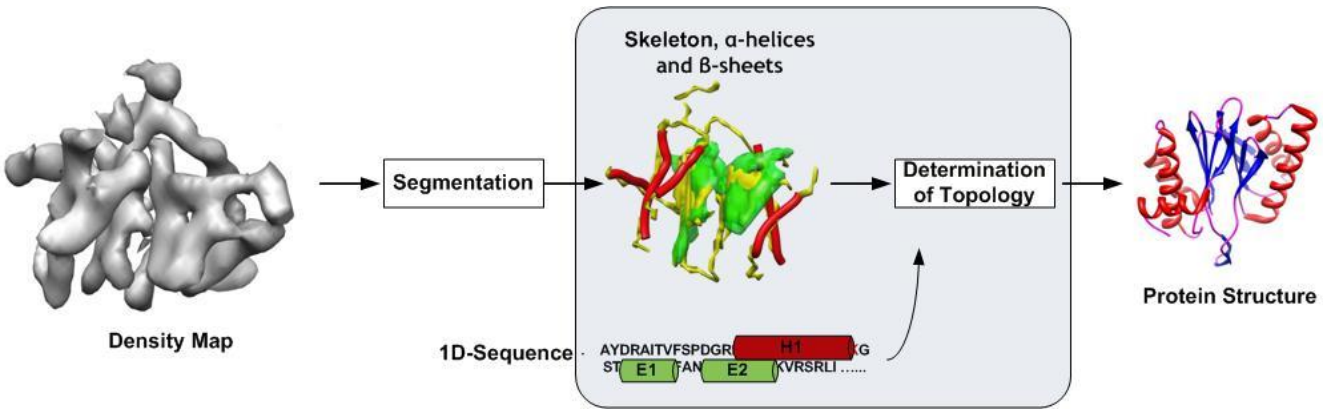


Fig. 1. Procedure of generating protein structure models from density map. Density map (gray) is a map segmented from 1733 (EMDB ID)[15]. Skeleton (yellow), α-helices (red) and β-sheets (green) are detected with density map[13], skeleton (yellow) of the density map is generated with Gorgon[16]. 1D sequence of the protein is a string, in which positions of α-helices (red) and β-strands (green) are highlighted. Atomic protein structure is the structure of protein 3C91(1733)

Present work focuses on the topology for β-sheets. More constraints for β-sheet topologies are involved to solidate the ability of screening unpreferred β-sheet topologies which were reported by Baker[18]. Same as the method used previously[7], a group of approximated parallel sticks are used to represent the β-sheet plane. The native β-sheet topology expected has top ranked path from graph (Fig. 2). However, unlike the α-helices, the connection traces between sticks are ambiguous. The skeleton obtained from density map in the β-sheet area is a twisted plance and can not provide reliable information for the topology. In current work, several constraints from the distribution of the β-sheet topologies are used to adjust the weight of the edges. Topologies with low occurrence probabilities have low probabilities to be the native topologies and will be screened from the candidates.

II. METHOD

Let $(H_1, H_2, .., H_{M_\alpha})$ and $(E_1, E_2, .., E_{M_\beta})$ be the SSEs of α-helices and β-strands predicted in sequence respectively, in which M_α and M_β are the corresponding number for each kind of SSEs. Let $(A_1, A_2, .., A_{N_\alpha})$ and $(B_1, B_2, .., B_{N_\beta})$ be the SSEs of α-helice sticks and β-strand sticks predicted on density map respectively, in which N_α and N_β are the corresponding number for each kind of SSEs. Without loss generality, assume $M_\alpha > N_\alpha$ and $M_\beta > N_\beta$. Each path in the graph represents a candidate topology for the native topology. All topologies can be enumerated from this $2*(M_\alpha+M_\beta)*(N_\alpha+N_\beta)$ 2D graph by involving direction of sticks. Each node is an assignment of SSEs in sequence to sticks and is set as (H_i, A_i, d) or (E_i, B_i, d), d is direction of sticks, 1 if sequence line passes the stick from the start point to the end point, -1 if take reverse direction. Weight for edge represents the possibility of having a loop between two sticks. Topology that only includes the assignment from E_i to B_j is called β-sheet topology. A simplified graph which only includes β-sheet topologies is shown in Fig. 2. Columns (B_1, B_2, B_3 and B_4) represent the four sticks that

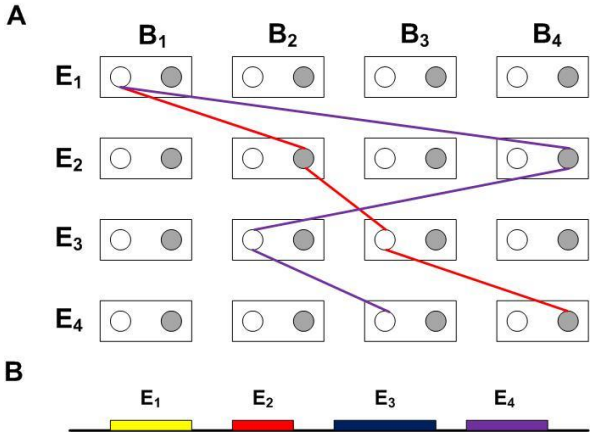


Fig. 2. β-sheet graph using in algorithm. A. coloumns B_1, B_2, B_3, B_4 represent the sticks of β-strands and rows E_1, E_2, E_3, E_4 represent the SSEs for β-sheet in sequence, white circle represents the parallel node that the direction of the sequence from N to C is identical with the direction of the stick points. Two different paths with different colors represent two different topologies. B. the SSEs predicted in the sequence, four consecutive β-strands in sequence are shown with different colors.

belong to the same β-sheet and rows (E_1, E_2, E_3 and E_4) represent the consecutive β-SSEs in the sequence (Fig. 2B). The connection relationship of (E_1, E_2, E_3 and E_4) SSEs in the sequence is shown in Fig. 2B. Let this chain go through the SSEs in the density map with the correct order and the correct direction. Define the positive direction of the sequence as the direction of protein chain from N to C, which represent amine (-NH2) and carbonyl group (C=O) in one amino acid respectively. Sticks representing the axis of α-helices and β-strands are detected with SSETracer[13]. Define the direction of sticks from the first point to the last point in the input file for each stick. In Fig. 3B and 3C, assume the stick directions for all four sticks are from the bottom to the top. Each stick and each

SSE in the sequence can be used only once. SSEs are assigned to sticks with either positive direction or negative direction and the same path in the graph can not pass the same stick twice. The sequence with different colors with respect to the strands in

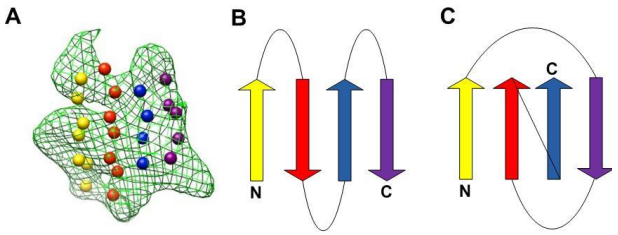


Figure 3. 4-Stranded β-Sheet. A: Density map in β-sheet area, four sticks with different colors represent the four strands in sheet; B: General topology; C: Rare topology

Fig. 3B and 3C is shown in Fig. 2B. True SSEs in PDB file are used here instead of predicted SSEs. Two paths are show in Fig. 2A which represent the topologies in Fig. 3B. and 3C. The topology corresponding to Fig. 3B has the order (E_1, B_1, 1), (E_2, B_2, -1), (E_3, B_3, 1), (E_4, B_4, -1).

Number of all candidate topologies is:

$$N = (C_{M_\alpha}^{M_\alpha - N_\alpha} \times N_\alpha! \times 2^{N_\alpha}) \times (C_{M_\beta}^{M_\beta - N_\beta} \times N_\beta! \times 2^{N_\beta}) \quad (1)$$

this huge searching space is simplified in He[17]'s algorithm that generates top K ranked topologies without enumerating all possible candidates. With an appropriate weight set for each edge, the native topology can be ranked within the top K topologies. Skeleton of density may is generated by Gorgon[16] with binary method and is used as the third input data. Density traces, d_{trace}, on the skeleton (yellow trace in Fig. 1) between two sticks with edge are measured and best fitting candidate to loop length (number of the residues between two SSEs in sequence times 3.8Å) is used to calculate the weight for edge:

$$W = d_{trace} - N_{loop} * 3.8 \quad (2)$$

Each path from the first row to the end row represents a candidate topology (Fig. 2A). In addition, occupancy filter is used to screen the inappropriate assignment. Maximum variation of length between SSEs in sequence and the sticks is 50%.

In addition to loop score and occupancy filter for α-helices, more constraints are designed for β -sheet topology. As shown in Fig. 3A, four sticks for β-strands are tiny twisted parallel lines with a distance ~4.5Å[19] between each other. This narrow gap may causes the loop score (2) is not sensitive enough to distinguish connections in β-sheet area. For example, topology in Fig. 3B is the most popular topology for 4-stranded β-sheet, which has three loops between strands. In the case that loop length is long enough (4 residues), loop score for the topology in Fig. 3B shows a low priority than the topology in Fig. 3C which is never observed in Dunbrack's database[20]. To solve this problem, several extra constraints are involved to reset the weight of the edge for β-sheet area:

A. Strand Spacing

Consecutive SSEs with a short loop (< 5 amino acids) prefer the antiparallel folding in the same sheet[18] as shown in Fig. 4A.

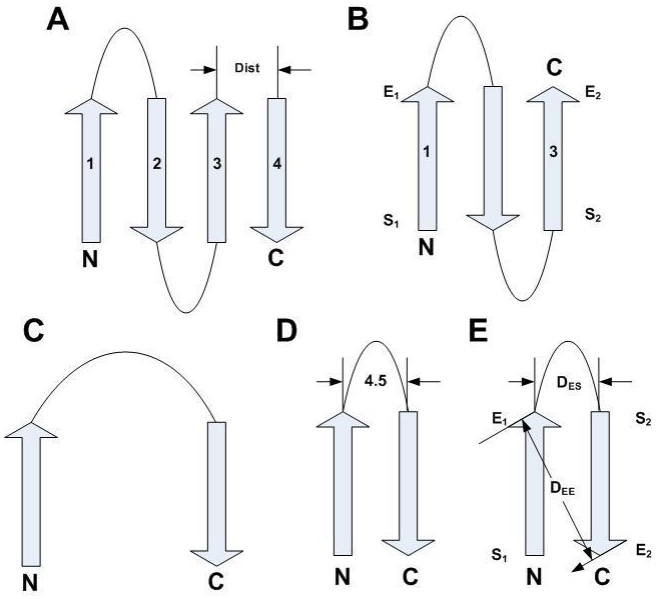


Fig. 4. Constrains for β-sheet topologies. A. Topology with highest occurrence probability, gap between strands is defined as the distance between two approximated parallel sticks. B. Define start point (S) and end point (E) for each strand. C. Greek-key topology. D. Neighbor SSEs in both sequence and density map. E. Define the distance between end point (E) of the strand and it's subsequent strand's start point(S), D_{ES} and the distance between corresponding end point (E) and end point (E) for neighbor SSEs.

However, topology in Fig. 3C whose gap between strands is large for consecutive SSEs in sequence is hard to be observed. Add a constraint to represent this space preference. Let $\text{gap}_{\text{seq}} = |i - j|$, in which i and j are the index of SSEs in sequence. Let $\text{gap}_{\text{stick}} = \text{dist}/4.5$, in which dist is the distance between two strand sticks. Set a penalty $50 * \text{gap}_{\text{stick}} - \text{gap}_{\text{seq}}$to the edge on the graph if two connected nodes have $\text{gap}_{\text{seq}} < \text{gap}_{\text{stick}}$. In Fig. 4A, gap between the first strand and the third strand is |3-1| = 2. If the edge connects strand 1 and strand 4 (Fig. 4C), this topology has very low occurrence probability since the loop between two consecutive SSEs in the sequence has to be long enough and the side-chains for this loop are difficult to be arranged without any overlap.

B. Strand Direction Penalty

Let the start point (S_1, S_2) to be the point in stick that protein chain goes in from N to C and the end point (E_1, E_2) to be the point comes out, $S_{1,2}$ and $E_{1,2}$ are points on sticks, D_{ES} and D_{EE} are distances between point E_1 on first stick and point S_2 on subsequent stick (Fig. 4E). Set a penalty to the edge if ($D_{ES} < D_{EE}$ and$mod(\text{gap}_{\text{seq}}, 2) = 0$) or if ($D_{ES} < D_{EE}$ and $mod(\text{gap}_{\text{seq}}, 2) = 1$) to improve the rank of the most popular topology in Fig. 4B. In most popular topologies, the chain enters the β-sheet from one side and leaves the β-sheet from the opposite side, then enters from leaving side with a high probability. Strand Direction Penalty adopts this entering-leaving character to improve the rank with the suitable

Table 1. Native Topology Rank in Top K Topologies

ID	#Helices[a]	#Sticks[b]	#Strand[c]	#Sticks[d]	Rank_C[e]	Rank_NC[f]
EMDB ID_5030	4	3	4	3	1	1
EMDB ID_1733	5	5	12	12	13	- / 100[g]
1OZ9	5	5	5	4	7	25
2KUM	2	2	3	3	1	5
2KZX	3	3	3	3	10	10
2L6M	2	2	3	3	6	6
1BJ7	5	1	9	9	4	- / 100
1ICX	6	3	7	7	1	2
1JL1	4	4	5	5	16	22

a. The number of α-helices in real protein structure
b. The number of α-helices predicted
c. The number of β-strands in real protein structure
d. The number of β-strands predicted
e. The rank of the native topology with the five constraints in present work
f. The rank of the native topology without the five constraints
g. -/100 means that the native topology is not within top 100 topologies

topology.

C. Neighbor Strands Reward

If the consecutive SSEs in the sequence are assigned to a pair of neighbor sticks (dist~4.5Å), the probability that the number of residues between two SSEs is less than or equals to 4 should be dominated (Fig. 4D). Set a reward for the edge which contains this kind of fundamental SSEs component. This constraint fixes the node pair assignments that are most obvious. A reward $-3.8 * \text{gap}_{\text{seq}}$ is added to this edge.

D. Node Direction Penalty

For the node pair which represents the consecutive SSEs in sequence and sticks are neighbor position in density map, D_{EE} should be greater than D_{ES} for this antiparallel strands pair. Parallel pair is inappropriate. Thus, set a penalty 150 for this edge if D_{EE} is less than D_{ES}. This constraint avoids the misleading from the loop score that sets low weight score for parallel topology.

E. Long Helix Matching

Based on the experience, long helices are more reliably predicted in both the sequence and the density map. Matching between long helices may be more accurate. If the length between the SSE in the sequence and the sticks on the same node is less than 15%, set up an extra reward -5 to this topology.

III. Results and Disscussion

Topology ranking algorithm in previous work[17] for the pure α-helices proteins has been enhanced to all proteins which may include both α-helices and β-sheet. Table 1 lists the results for some nine proteins. EMDB 5030 and EMDB 1733 used the real density maps which are measured with CryoEM. Both density maps and the solved PDB files for them are downloaded from EMDB. The other seven samples used the PDB files downloaded from PDB bank and corresponding density maps are simulated with resolution 8 Å by Chimera. 1BJ7, 1ICX and 1JL1 are samples used in EM-Fold[4]. Similarity of the samples in the sequence is less than 20% to avoid homologous samples involved. 2nd,and 3rd columns in Table 1 represent the number of α-helices in real protein structures and the number of sticks predicted. 4th and 5th columns are same number for β-sheet. 6th column includes the rank of the true topology after constraints are added. 7th column lists the rank of the true topology without any β-sheet constraint.

The ranks for all samples have been improved with β-sheet constraints. 1733 has total 17 sticks (5 helical sticks and 12 strand sticks) and the native topology is not found within top 100 topologies when only loop score and occupancy constraints. The major trouble is from the 2-stranded β-sheet whose 2 strands are tight close to other two 5-stranded β-sheet. This packed area makes the loop score lose the ability to recognize which sticks belong to the same sheet. Sticks in this 2-straned β-sheet trend to be treated as part of other two β-sheet due to the loop is long enough to connect several sticks from different β-sheet area even the loop residues number may only 4 or 5. Strand Spacing limits that strand connect to a far strand in same β-sheet or different β-sheet; Strand Direction Penalty reduces the possibility to form a parallel topology; Neighbor Strand Reward improves the priority to connect the neighbor strands and reduces the rank of the topology that includes jumps. Node Direction Penalty fixes the antiparallel topology or the parallel topology for the neighbor strands and decreases the rank of the wrong topology (antiparallel or parallel). Long Helix Matching aligns the longest α-helix (LEU48-ARG71) that reduced the α-helices searching space. Rank for 1BJ7 is improved to top 1 from a rank worse than 100. In EM-Fold, 1BJ7 is not

recognized in assembly step due to the weak β-sheet constraints. These two samples may display how necessary to add β-sheet constraints to the algorithm. Other seven samples have the rank within 100 even no constraints are applied. This may be due to the loose arrangement of their sticks. Sticks for β-sheet in these samples can be treated as the loops without significant negative effect.

Density map is ambiguous due to the low resolution of samples. This causes that not all sticks can be predicted from density map. 5030, 1BJ7 and 1ICX has missed sticks which are not recognized from the density map. This increases the degree of freedom of assigning a SSE in the sequence to a stick. However, ranks in Table 1 for these three proteins have no significant difference from other results. This may show that algorithm is resistant to stick prediction missing.

The native topology of a protein with β-sheet is sensitive and difficult to be distinguished. β-sheet has a lower percentage of secondary structures and a higher percentage of loop regions[6]. Flexible loops with high degree of freedom make the connection very complex between strands within the same sheet or among different sheets. Loop score in previous algorithm prefers the connection between two edge strands and may choose the unexpected topologies as the top topologies. By involving some of the β-sheet characters, the unpreferred topologies are screened and set the native topology within the top 20 topologies. More constraints may further improve the algorithm.

IV. Conclusion

Applying the extra constraints that are extracted by analysis of protein structures in PDB bank improves the recognization ability for the native topology. In present work, only the edge weight for pairs of nodes are reweighted which can cover the local characters. Global distribution of β-sheet topologies has been studied in Ma's paper[6] which can be involved in our future work. Furthermore, the statistics analysis information of SSEs organization in both the sequence and the 3D-space can be used to optimize the global score of the native topology.

References

[1] B. Bottcher, S. A. Wynne and R. A. Crowther, "Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy," *Nature, 386, 88-91,* 1997.

[2] S. J. Ludtke, D. H. Chen, J. L. Song, D. T. Chuang and W. Chiu, "Seeing GroEL at 6Å resolution by single particle electron cryomicroscopy," *Structure, 12, 1129-1136,* 2004.

[3] M. Billeter, "Compaison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals," *Q Rev. Biophys, 25, 325-377,* 1992.

[4] S. Lindert, R. Staritzbichler, N. Wötzel, M. Karakas, P. L. Stewart and J. Meiler, "EM-Fold: De Novo folding of α-helical proteins guided by intermediate-resolution electron microscopy density maps," *Structure, 17, 990-1003,* 2009.

[5] S. Lindert, N. Alexander, N. Wötzel, M. Karatkas, P. L. Stewart and J. Meiler, "EM-Flod: De Novo atomic-detail protein structure determination from medium-resolution density maps," *Structure, 20, 464-478,* 2012.

[6] Y. Wu, M. Chen, M. Lu, Q. Wang and J. Ma, "Determing protein topology from skeletons of secondary structures, " *J. Mol. Biol., 350, 571-586,* 2005.

[7] Y. Lu, J. He and C. M. Strauss, "Deriving protein structure topology from the helix skeleton in low resolution density map using ROSETTA, " *J. Bioinform. Comput. Biol., 6(1), 183-201,* 2008.

[8] F. Birzele, S. Kramer, "A new representation from protein secondary structure prediction based on frequent patterns," *Bioinformatics, 22, 2628-2634,* 2006.

[9] L. J. McGuffin, K. Bryson, D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics, 16, 404-405,* 2000.

[10] G. Pollastri, D.. Przybylski, B. Rost and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, " *Proteins, 47, 228-235,* 2002.

[11] G. Pollastri, A. McLysaght, "Porter: a new, accurate server for protein secondary structure prediction, " *Bioinformatics, 21(8), 1719-1720,* 2005.

[12] 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., 308, 1033-1044,* 2001.

[13] D. Si, K. Al Nasr, J. He, "A machine learning approach for the identification of protein secondary structure elements from electron cryo-microscopy density maps", *Biopolymer, 97(9), 698-708,* 2012.

[14] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The protein data bank: a computer-based archival file for macromolecular structures," *J. Mol. Biol. 112, 535-542,* 1977.

[15] M. Tagari, R. Newman, M. Chagoyen, J. M. Carazo, K. Henrick, "New electron microscopy database and deposition system," *Trends. Biochem. Sci., 27(11), 589,* 2002.

[16] T. Ju, M. Baker, W. Chiu, "Computing a family of skeletons of volumetric models for shape description," *Comput. Aided Des., 39(5), 352-360,* 2007.

[17] K. Al Nasr, D. Ranjan, M. Zubair and J. He, "Ranking valid topologies of the secondary structure elements using a constraint graph," *J. Bio. Comput. Bio., 9(3),* 415-430, 2011.

[18] I. Ruczinski, C. Kooperberg, R. Bonneau and D. Baker, "Distributions of beta sheets in proteins with application to structure prediction," *Proteins: Structure, Function, and Genetics, 48, 85-97,* 2002.

[19] T. C. Terwilliger, "Rapid model building of β-sheets in electron-density maps," *Act Cryst., D66, 276-284,* 2010.

[20] G. Wang, R. L. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics, 19, 1589-1591,* 2003.

Gaming and Virtual Reality

VMASC Track Chair: Mr. Hector Garcia
MSVE Track Chair: Dr. Yuzhong Shen

Natural User Interfaces for Learning and Exploration
Author(s): Gary Lawson, Rifat Aras, and Yuzhong Shen

Lessons Learned from iOS Application Development: Towards Apps for M&S and Games
Author(s): Anitam Das Nirjhar

Use of Interactive Immersive Visualization for the Control of Dental Anxiety During Dental Hygiene Treatment
Author(s): Carmelo Pardino-Barrios, Gayle McCombs, Norou Diawara, and Gianluca De Leo

Natural User Interfaces for Learning and Exploration

Gary Lawson, Rifat Aras, and Yuzhong Shen

***Abstract*—Natural User Interface is an emerging Human-Computer Interaction technique that allows creation of natural, intuitive, and easy to use applications. In the presented work, two applications with natural user interfaces are demonstrated with emphasis on learning and exploration using the Microsoft Kinect. The efficiency of the natural user interfaces may enhance the learning outcome and augment the sense of exploration in a virtual environment.**

I. Introduction

A. Brief History of Human-Computer Interaction

USER interfaces (UI) for computing have come a long way since the beginning of the computing. In 1946, the ENIAC was programmed using telephone exchange style re-wiring. Then in 1960's, the internal architecture of the computer was exposed in the UI controls through the front panel switches. Using the switches, the computer was programmed one word at a time. After the reign of switches, the revolution of batch processing brought the punch cards, many of which were punched at a time, handed over to an operator, who put them in line with everyone else's cards. After your cards were processed, which can take hours, you could pick up your cards and hopefully along with them your output. Punch cards were followed by operator consoles, remote terminals, and command line interfaces (CLI) that are still in use today. The debut of the mouse as an input device was in 1968 by Engelbart et al. [1]. The famous demo of Engelbart's team in 1968 introduced not only the mouse, but also other important concepts like hypertext, dynamic file linking, and even desktop sharing through a network connection. The term WIMP (windows, icons, menus, pointer) was brought up by Alan Kay in 1973 and the term was coined by Merzouga Wilberts in 1980 [2]. The WIMP term was an important stepping stone in describing building blocks of the modern user interfaces and this paradigm still dominates the graphical user interface design decisions in the mainstream computer applications.

B. Natural User Interfaces

The paradigm shift from CLI to WIMP improved the usability of the systems and reduced the learning time of every-day consumers from months to weeks. The next evolutionary phase shift in the user interface design is considered to be the introduction of natural user interfaces [3]. Natural user interfaces (NUI) (also known as direct user interfaces and metaphor-free computing) try to drive the user interface design from abstract concepts to natural representations. Instead of getting in front of the application, the interface based on natural elements disappear to the user; they fade to the background, and this results in quick learning, higher efficiency, and effortless usage.

Because of the appeal of the subject and the idea of more efficient user interfaces, researchers have been experimenting with natural user interfaces since 1970s. In 2006, Jefferson Han presented his multitouch system "Perceptive Pixel" that was based on his work on frustrated total internal reflection displays [4]. In this work, he introduced metaphor-free interaction techniques such as "pinching" gesture to separate a highly viscous fluid into two parts. Using a traditional user interface, this operation would require first selecting a metaphor for the corresponding tool and then applying this tool to the correct location of the object.

After its launch in late 2010, Microsoft Kinect received attention not only from the gamers but also from NUI developers as well. Microsoft provides its own implementation of an NUI for use with the Kinect hardware. The NUI API provides access to the raw color, depth, and audio feeds as well as the processed skeleton joint locations in physical space and spoken grammars recognized by the voice recognition software. This provides seamless interaction between users and the Kinect-based application with minimal effort, further promoting intuitive interfacing. Recently, the NUI API has been extended to include gesture support similar to those found with multi-touch displays. Swiping, pinching, spreading, waving, and pushing are now common gestures which may be utilized with the Kinect and aid in the intuitive nature of NUI's which is the underlying goal for human-Kinect interaction.

The next section, *Application with Natural User Interfaces*, is segmented into two major components. Each component discusses an application that was developed using various NUI techniques. The first application is a virtual environment which allows users to explore a visualization lab at Old Dominion University. The second is Tangible Table Top; an application developed to facilitate collaborative learning and exploration. Both applications incorporate the Kinect which provides multimodal interaction schemes unique to each.

II. APPLICATIONS WITH NATURAL USER INTERFACES

A. Virtual Environment and Web-Space Exploration with an NUI

In the human-computer interaction context, modalities are the different means of sending/receiving information to/from the computer. In a multimodal interaction scheme, any human sense can be a method of transferring information, the major ones being vision, audition, tactition, and proprioception. Providing multiple modalities to a user in a balanced way can increase the usability of the system and also augment the cognitive processes. In natural user interface design, multimodal interaction schemes can be employed to include more natural communication styles to the interface, such as adding natural language voice recognition to a CAD/CAM application.

A Kinect-based application is a prime example of a multimodal system as it takes advantage of vision, audition, and proprioception and use these senses to perform a task. This is important in virtual environments as the purpose is to immerse a user into an environment and providing the sense of exploration. Multimodal systems can augment the immersion of a user if the multiple senses are utilized in a balanced way that can be achieved by accurate design of the user interface elements.

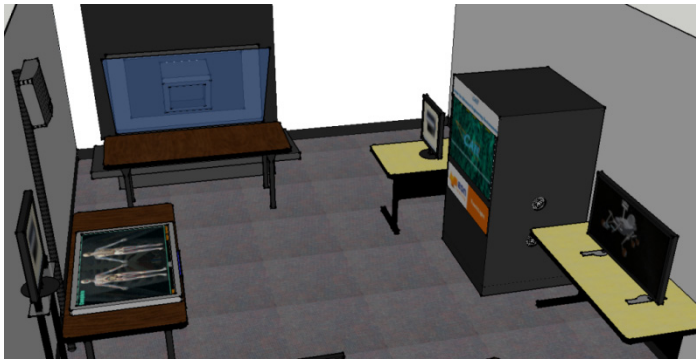


Fig. 1. Distributed Visualization Lab at ODU's Pensinsula Center. The stations follow: tabletop display (bottom left), Ipresence video conferencing table (top left), Touchlight display (top right), autostereoscopic display (bottom right).

The virtual environment in this work is the Distributed Visualization Lab (DVL) located at ODU's Peninsula Center. It is a well-spring of communication and visualization technologies that hosts several visualization stations which feature very different methods of interaction. Within the lab there exist four devices which have been duplicated in this work as shown in Figure 1: the Touchlight display, autostereoscopic display, the table top display, and the Ipresence video conferencing table [5]. The Touchlight device is a visualization tool with the combined benefit of gesture-based multimodal interaction. It is parallel to that of the Kinect device; except it also handles visualization and interfacing. The autostereoscopic display is used for rendering three-dimensional images that can be viewed without the use of special headgear or glasses. Ipresence is a teleimmersive device which combines tele-presence and eye-to-eye contact over distances. The tabletop display is a multi-user, debris-tolerant device which responds to touch and gesture commands. Users may engage the device by leaning up against a mat which gives the user an electrical charge. Touching the surface is interpreted similar to multi-touch displays commonly used in cellular devices. Each of these devices can be considered as a precursor to the devices commonly used today.

This work introduces a Kinect-based virtual environment application which has been developed to explore the lab. There are two final implementations for the virtual environment, a web-based export and a stand-alone application. The web-based export adopts traditional mouse and keyboard interaction techniques and provides users with functionalities that allow for quick and intuitive navigation of the environment. The stand-alone application is Kinect-based and incorporates static handling and gesture techniques to also allow for quick and intuitive navigation.

Interaction Techniques

Providing users with a seamless NUI is the goal in modern applications. They should be intuitive, complete, and easy to learn and remember when not intuitive. Therefore this work implements interaction techniques from traditional mouse interaction, gesture input, and speech recognition for use with the Kinect device. From the traditional mouse interaction aspect, the user's hand acts as a cursor. As the user's hand is waved in physical space, the cursor on screen reflects the movement. Hovering over an object is used for selection no matter if the user is expected to select a button or a station. Additionally, the user may use the boundaries of the screen to adjust the viewing angle within the environment. This method is simplistic and highly accurate as the user may finely adjust the viewing angle or hastily rotate around to see objects behind them. However, this method is tedious in nature and does not provide a very immersive method of interaction.

Users may choose to navigate using gestures. On a mobile device, it's common to rotate an object by swiping or gradually sliding a finger across the screen. The prior offers a quick and easy method for the user to rotate the object while the latter offers the user a high precision technique for rotating the object to just the right angle. The interaction scheme in this application follows the same general idea. There are difficulties with gestures and the Kinect however. With touch screens, a gesture cannot be detected until a user touches the screen; and then any gesture to be captured waits for a specific set of actions made by the user. The Kinect does not have an on/off state for gestures and therefore all actions made by the user must be considered when detecting gestures. However, with the release of later versions of the Kinect SDK, more refined gesture techniques have been developed allowing for more accurate gesture capturing. Thusly, instead of strictly forcing users to move based on holding their hand to the side of the screen; users may perform a swipe which will rotate the camera. The speed and direction of the rotation are determined by the speed and direction of the swipe allowing users to intuitively rotate the camera. Due to complications with detecting user's intents with the Kinect, the option to rotate by holding a hand at the side of the screen is still implemented as

it provides users with another, more accurate method of rotating the camera.

Another interaction technique must be implemented however to accommodate the use of gestures, the resting state. The resting state is important to recognize and allow in an application such as this because the user's hand position directly manipulates the screen. Without requiring more specific movements from the user, such as holding their hand a specific distance from their body, it is difficult to determine when the user would like to rotate the camera and not; again, there is not an easy method of turning on/off gestures with the Kinect. To counter this problem, a rest condition is implemented allowing users to rest their hands at their side. When this condition is detected, the application halts updating hand positions until the user places a hand back into the active area (above the waist).

Users may speak commands that the Kinect will detect, recognize, and respond to. In this application, users may voice different commands depending on the current state of the application. Table 1 provides a listing of all the voice commands for each state and the response of the Kinect when the command is recognized. During the station selection mode, users rotate the camera around the room to determine a station to select. They may select one using their hand and hover over the station they wish to select, or voice the command "Station <#>". Upon selecting a station, a different UI appears offering the users various options for exploring the station. For the Touchlight device and autostereoscopic display, users may only view a set of images which may be shown on the physical device. These are the most simplified stations. The I-presence device allows users to watch a short video about immersive classroom environments. Finally the table top display allows users to interact with the device by selecting from a list of 4 different human systems: skeletal, nervous, circulatory, and organs. Once selected, a view of the system within the human body is shown; again something which may be viewed on the device itself. Users may then transition back to the station selection state and select a new device to explore by voicing the command "back".

TABLE I
VOICE RECOGNITION COMMANDS AND APPLICATION RESPONSES

Application State	Voice Commands	Response
Start Menu	Start	Play the introductory animation.
Station Selection	Help	Bring up the help menu. Provides gesture listing, performance examples, and voice command listing.
	Station <#>	Transition to station #: 1. Touchlight display 2. Autostereoscopic display 3. Table top display 4. I-Presence video conferencing table
	Backward Behind Reverse	Rotate the camera 180 degrees.
Station Exploration	Help	Bring up the help menu.
	Next Previous	View the next or previous image (stations 1 and 2)
	View <#>	View the human system #: 1. Skeletal 2. Nervous 3. Circulatory 4. Organs (station 3)
	Play Pause	Play or pause the video (station 4)
	Back	Transition to station selection

B. Tangible Table Top: Collaborative Learning and Exploration Environment

A tangible user interface is a class of natural user interface, in which the user interacts with the digital information through manipulating physical objects. In order not to distort the natural interaction with the digital information, physical objects have to be perceptually coupled in a natural way to the actively manipulated digital information.

In the early work of Underkoffler and Ishii [6], the authors presented a tangible workbench concept for urban planning and design. By combining physical building models and virtual interaction tools, they were able to assess and find solutions to several urban planning problems such as preventing buildings casting shadows at each other at different times of day, resolving reflected light on the neighboring highway, and decreasing the effect of pedestrian-level windflow. Underkoffler and Ishii called this system the luminous-tangible interaction, which consists of manipulation of physical objects and seeing the effects of the changes as an ongoing projection of visual information.

A more recent work about the subject applies the same workbench approach to support disaster education [7]. The user can interact with the system through manipulating physical objects attached with LC tags. For example, a user can assess an evacuation strategy by placing a physical house model on the workbench denoting the start location of the evacuation. Likewise, the evacuation route can be input to the system by using a digital pen as the input device.

The presented work, Tangible Table Top, applies the discussed tangible workbench concepts on collaborative learning and exploration tasks. It features multitouch interaction without the need of expensive custom hardware, enhances collaborative nature of the environment by allowing multitouch events from mobile devices, and recognizes and tracks physical objects to show spatially correct contextual information to improve learning outcomes. The low-cost hardware and open-source software components of Tangible Table Top makes it particularly appealing for a large user-base.

1) Hardware Used

The hardware setup of Tangible Table Top consists of a projector and a Microsoft Kinect device arranged in a vertical orientation (Figure 2). The data received from the Kinect device is processed by a mainstream personal computer and the visual output is sent to the projector to be displayed on any surface.

Fig. 2. The conceptual design of the hardware setup of the Tangible Table Top is depicted. The Kinect device and the projector are arranged in a vertical orientation looking downward towards a surface.

The hardware requirements of the Tangible Table Top are basic, low-cost ones, and this increases the application potential of the system greatly.

2) Software Used

The software that drives the operations of Tangible Table Top is arranged in a modular structure. The modules are separated according to their functionalities and implemented using open-source libraries (Figure 3).

The data flow of the application is initiated from the Kinect Object component. This component provides a single point of access to the Kinect device through the Microsoft Kinect SDK and streams the depth and RGB images from the device to the OpenCV Object component. This component, as its name implies, uses the OpenCV library [8] to process the received depth and color images, extract meaningful information, and forward the information to the appropriate points in the data flow. Specifically, two facilities are developed in the OpenCV Object component to capture the multitouch points (Blob Tracker) and track the physical object (AR Tracker).

The detected multitouch cursor candidates are sent to the TUIO Listener component. TUIO is an open framework that defines a common protocol to be consumed by tangible user interface implementations [9]. TUIO is a device independent protocol and it is adopted widely by the NUI community. Several frameworks have been implemented using this protocol allowing integrating various devices into one's own application hassle-free. The TUIO Server component works in conjunction with the TUIO Listener component and forwards all of the multitouch events to the OSG Object component, which is based on the well-known OpenSceneGraph library [10]. The OSG Object component along with handling the multitouch events, also utilizes the location and orientation information of the recognized physical object to visualize spatially correct contextual information.

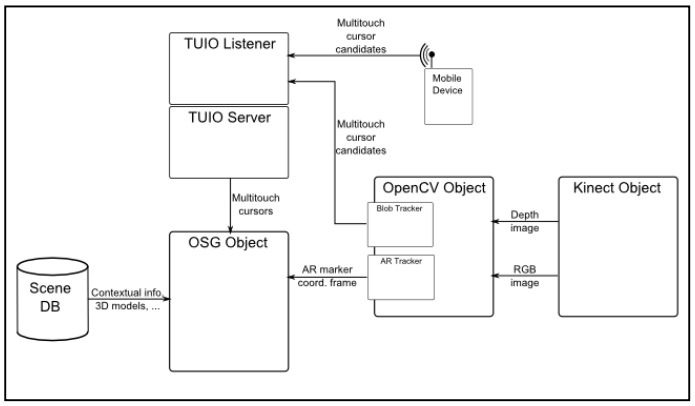


Fig. 3. The component diagram of the Tangible Table Top application is shown.

These components work in symphony to create a tangible user interface experience that mixes the physical representation with virtual information to improve the learning outcome in a collaborative explorative virtual environment (Figure 4).

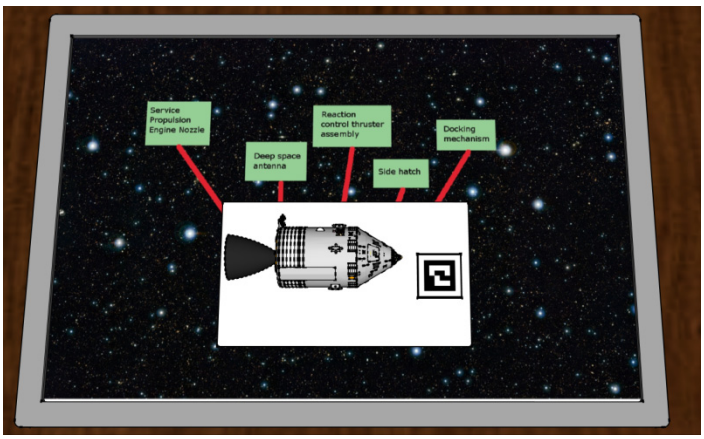


Fig. 4. The 3D printed model of the Apollo CSM is placed on the Tangible Table Top surface. The physical model is recognized by the TTT, and the scene background is changed to an appropriate one. Spatially correct contextual information is presented to the user.

III. Conclusion

The paradigm shift from punch cards, batch processing, and command line interfaces to the WIMP model had a great impact on the usability and efficiency of the user interfaces. The shift from WIMP model to natural user interfaces is expected to have a greater impact, essentially because NUIs allow users to interact with the system in a natural intuitive way as the interface disappears to the user. Natural user interface research is an ongoing one, and this new interaction technique is to be applied to other disciplines such as modeling and simulation. In this paper, two applications with an emphasis on learning and exploration are presented with natural user interfaces. The design principle behind the application interaction is to be natural, intuitive, and easy to use.

REFERENCES

[1] D. C. Engelbart and W. K. English, "A research center for augmenting human intellect," in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 1968, pp. 395-410.

[2] B. A. Myers, "A brief history of human-computer interaction technology," *interactions,* vol. 5, pp. 44-54, 1998.

[3] A. de los Reyes, "Predicting the Past," *Web Directions South 2008,* 2008.

[4] J. Y. Han, "Low-cost multi-touch sensing through frustrated total internal reflection," in *Proceedings of the 18th annual ACM symposium on User interface software and technology*, 2005, pp. 115-118.

[5] O. D. University. (2013). *Intelligent Collaborative Visualization Hub*. Available: http://www.aee.odu.edu/facilities_visualization.php

[6] J. Underkoffler and H. Ishii, "Urp: a luminous-tangible workbench for urban planning and design," in *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, 1999, pp. 386-393.

[7] K. Kobayashi, T. Kakizaki, A. Narita, M. Hirano, and I. Kase, "Tangible user interface for supporting disaster education," in *ACM SIGGRAPH 2007 posters*, 2007, p. 144.

[8] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*: O'Reilly Media, Incorporated, 2008.

[9] M. Kaltenbrunner and R. Bencina, "reacTIVision: a computer-vision framework for table-based tangible interaction," in *Proceedings of the 1st international conference on Tangible and embedded interaction*, 2007, pp. 69-74.

[10] R. Wang and X. Qian, *Openscenegraph 3.0: Beginner's Guide*: Packt Pub Limited, 2010.

Lessons Learned from iOS application development: Towards apps for M&S and Games

Anitam Das Nirjhar, Computer Science Department, Old Dominion University

Introduction

In recent years, iOS based devices have been getting popular. According to polls, sales of iOS based devices has been gone up to 250 million units. This makes iOS a widely available platform for running models,simulations and visualizations. This paper reports on lessons learned, challenges and insight, on how an iOS based application is built in order to learn how to build and run iOS based games.

Motivation

Making both 3D and 2D games or applications is a great choice for programmers because they reach a wider audience. Reaching a wider audience means more revenue and if the goal of the applications is educational, it can make a global change. Until recently, the capability of running 3D games was reserved to computers. Today, smartphones have that capability. iOS based devices provide a lot functionality and tools which can handle 3D and 2D graphics in a fast and smart way. In order to make 3D or 2D games and applications for iOS devices one has to know the procedures and techniques of programming for the iOS based operating system as iOS devices have been proved themselves appropriate.

Four main reasons have been identified why iOS based devices are considered to be appropriate for an application.

Wide audience: iOS based devices are getting more attention more over the years. Apple has 357 stores in 11 countries. There are 60 millions Mac users around the world. Apple has sold 45 millions iPods so far this year. There are more than 500,000 apps in the app store, 140,000 of them are iPad apps.

Solid operating system(iOS): iOS operating system provides a much more user friendly environment. The built in applications makes the device more reliable in common uses. The power of Multi-tasking and Notification Center creates a whole new user experience never seen before.

Advanced Hardware: iOS based devices comes up with better hardware in every launch of a new device. Most of them has Fast processors, Retina Displays and various Sensors, Multi touch interface, three-axis gyro. Because Apple makes both the hardware and the operating system for iPad, iPhone, and iPod touch, everything is designed to work together.

AppStore: Once an application is ready for launch, it can be uploaded to App Store by Apple, which provides excellent services for selling and distributing applications. Applications needs no extra hosting sites as App Store provides their own hosting for every application. If user purchases any application, developers can easily get the payment through App Store as it maintains every transaction itself.

Despite these advantages, programming games in the iOS platform is not without challenges.

Challenges

To build 2D and 3D applications in the iOS environment one has to meet a set of challenges.

Friday, March 2013.
Anitam Das Nirjhar
Computer Science Department
Old Dominion University
anirj001@odu.edu

Getting used to MacOSX: In order to build an application, one has to get used to Mac operating system. Recently there are some tools available, which can help you build iOS based applications in Windows. But, the programmer might not get the full usage of SDK provided by Apple. The best procedure would be building the application in Mac operating system. Programers have to be acquainted with the basic structure of Mac operating system as they will have to deal with various formats of both system and project files.

Programming in Objective C: The native language for iOS is Objective C and it has a different syntax style comparing to other Object Oriented Programming languages. It doesn't follow the conventional way of object oriented programming as seen in Java or C#. Learning Objective C is a very hard task. Even for starting working in a project, one has to have the working knowledge of Objective C. There are several resources in the internet to learn this language.

Getting used to the IDE: The only IDE that is used to make an iOS based application in Mac environment is Xcode. This IDE is user friendly enough to help you manage your project files. Xcode creates a virtual file structure for your project and binds the files to the system automatically when you run your project. Using Xcode you can easily change your project settings anytime and manage emulator's rotational behavior. Working in Xcode maybe a little tough in the first times, but user will get used to it in a while.

Working with GLTK: In an application, graphics is really an important part. Using OpenGL, programmer can easily build both 2D and 3D views of multiple objects. GLTK is a viewer which can render OpenGL code in iOS devices and show 2D or 3D objects in an application. Working in OpenGL is a very interesting task.

Getting information from Sensors: iOS devices have multiple sensors which can be used as a source of data. Several applications have already been made that completes their tasks using data gathered from sensors. Most of the iOS devices have Gyroscope which can provide 3D position of the device. A lot of games and applications have been made using gyroscopic data.

Developing Universal Applications: Making an universal application is really a hard task. One application should be made in a such a procedure, which will run in all kinds of iOS devices. Different kind of iOS devices have different screen sizes and resolution. So, automation of screen size and resolution is required in an application to make it universal.

Insights

Throughout the whole process of making an application for iOS device, user will learn two main lessons: learn about the programming environment and also importantly about basic project management skills. The concept of programming in iOS environment operating system will be clear to the user as he/she will be working with the native language Objective C. The intercommunication between the application and the system will be far more interactive to the user. Also, a user will know how to divide a whole project in individual procedures and work in a schedule. While working, user shall get acquainted with more ideas to make an application more user friendly and efficient. There can be problems while making an universal application, but user has to look for alternative solutions that wont change the main plan of the project.

Conclusion

This paper illustrates the lessons learned from iOS development towards M&S and Games. These lessons learned are presented in the form of challenges and insights. Besides learning about the iOS environment, the author also found the importance of proper project management.

USE OF INTERACTIVE IMMERSIVE VISUALIZATION FOR THE CONTROL OF DENTAL ANXIETY DURING DENTAL HYGIENE TREATMENT

Carmelo Padrino-Barrios, RDH, BSDH, Gayle McCombs, RDH, MS, Norou Diawara, PhD, Gianluca De Leo, PhD, MBA.

I. INTRODUCTION

Dental anxiety has been reported as one of the barriers individuals face when seeking dental care [1]. Dental anxiety is defined as a nonspecific nervousness, tension, or concern about any preventive or restorative dental procedure [1]. In fact, dental fear is categorized as one of the most common fears along with flying, heights, and closed spaces [2]. Unfortunately, dental anxiety is considered a common response in dental offices. Poor oral hygiene and inconsistent dental visits are often related to dental anxiety and pain. People can develop anxiety at any stage of their lives, but child-onset shows a more severe negative response than adolescence or adults [3].

When managing the anxious dental patient, the American Dental Association recommends the use of pharmacological agents to make the experience as comfortable as possible [4]. Although pharmaceutical medications can help clients with anxiety, the efficacy of these treatments is not guaranteed. Therefore, new techniques such as the use of audiovisual immersion therapy and virtual reality (VR) are being studied to assist anxious patients. VR uses artificial or computer-generated sensory experiences, and are considered an alternative technique to control anxiety and fear episodes [5,6]. Researches suggest that VR is a viable option due to its effectives to minimize anxiety and phobia symptoms [5,6,7,8,9].

The use of audiovisual eyewear helps clients obtain a more positive experience in the dental office. Research indicates that VR and audiovisual distraction have a positive effect in the reduction of pain [9]. VR can be applied in any medical environment through the use of intra-operative video [7].

Many health care professional and consumers want to move away from pharmaceuticals and Interactive immersive visualization (IIV) provides an alternative or complimentary treatment to manage dental anxiety. The purpose of this study is to measure the effects of IIV and the level of anxiety in patients during a dental hygiene treatment.

II. METHODS

A convenience sample of 30 participants will be chosen from the Hampton Roads area. A split mouth design will be utilized. All patients will have their whole mouth cleaned, starting with the right side, but will be randomized to the following:

- ½ mouth cleaned + IIV
- ½ mouth cleaned without IIV

All the appointments will be performed in the Dental Hygiene Research Center (DHRC), room #1101 in the Health Science Building at Old Dominion University, Norfolk, Virginia. A registered dental hygienist will perform all procedures associated with hygiene treatment. The dental hygiene examiner will determine the potential eligibility by conducting a pre-screening phone questionnaire. Medical and dental histories and Cohran's Dental Anxiety Scale (DAS-R) will be performed to determine health status and anxiety level. The dental hygiene treatment consists of a whole mouth

nonsurgical periodontal procedures (dental cleaning).

The inclusion criteria in the study is to be: generally healthy, adults males or females 18 years or older, able to understand the purpose of the research, sign consent, consent to wear the IIV headset, have a baseline score of 9 or higher on the Corah's Dental Anxiety Scale (DAS-R), have no severe dental caries, calculus, or periodontal disease, take no medications for anxiety or required antibiotic premedication.

Subjects will wear the IIV eyewear as an audiovisual tool during the appointment. Each subject will have the opportunity to select between three different videos: a documentary video, a series of music videos, or a TV show with accompanying audio. Instructions for the use of both interactive immersive eyewear will be provided prior to use before treatment begins. In addition, subjects will be able to adjust the volume settings in the interactive immersive eyewear.

Subjects will complete the DAS-R scale as a pre-screening to determine eligibility and baseline. A Calmness scale will be conducted pre-, during and post-treatment. At the completion of the study a IIV opinion survey will be conducted to determine how much comfort the eyewear provided to the subjects.

III. RESULTS

The current study has been collecting data since the end of the fall of 2012. Currently eighteen subjects have completed the study. The preliminary results have indicated that the use of IIV reduces the anxiety level in subjects during dental hygiene treatment. However, a statistic analysis needs to be implemented prior to determine the effectiveness of IIV on dental anxiety during dental hygiene treatment.

IV. CONCLUSION

Subjects can experience anxiety when seeking dental care. The use of interactive immersive visualization could promote a calmer state to the patient during dental hygiene treatment. Interactive immersive is an effective and less expensive technique that can be implemented at any dental site that offers dental hygiene treatment.

REFERENCE

[1] Darby, M. L, & Walsh, M.M. (2003). Behavioral management of pain and anxiety. *Dental hygiene theory and practice* (pp. 672-684). St. Louis Missouri: Saunders.

[2] Fiset, L., Milgram, P., Weinstein, P., & Melnick, S. (1989). Common fears and their relationship to dental fear and utilization of the dentist. *American Dental Society of Anesthesiology. 36*, 258-264.

[3] Locker, D., Liddell, A., Dempster, L., & Shapiro, D. (1999). Age of onset of dental anxiety. *Journal of dental research*, *78*, 790-796 doi: 10.1177/00220345990780031201

[4] American Dental Association, (2007). Policy statement: the use of sedation and general anesthesia by dentists. http://www.ada.org/sections/about/pdfs/statements_anesthesia.pdf

[5] Maltby, N., Kirsch, I., Mayers, M., & Allen, G. (2002). Virtual reality exposure therapy for the treatment of fear of flying: a controlled investigation. *Journal of Counseling and Clinical Psychology*, *70*(5), 1112-1118.

[6] Parsons, T., & Rizzo, A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of Behavior Therapy & Experimental Psychiatry, 39*(3), 250-261.

[7] Man, A., Yap, J., Kwan, S., Suen, K.L., Yip, H.S., Chen, P.P. (2003). The effect of intra-operative video on patient anxiety. *Anaesthesia, 58*(1), 64-68. doi:10.1046/j.1365-2044.2003.02788 4.x

[8] Ram, D., Shapira, J., Holan, G., Magora, F., Cohen, S., & Davidovich, E. (2010). Audiovisual video eyeglass distraction during dental treatment in children. *Quintessence International, 41*(8), 673-679.

[9] Wismeijer, A., & Vingerhoets, A. (2005). The use of virtual reality and audiovisual eyeglass systems as adjunct analgesic techniques: A review of the literature. *Society of Behavioral Medicine*, *30*(3), 268-278.

Training and Education

VMASC Track Chair: Mr. Sol Sherfey
MSVE Track Chair: Dr. Roland Mielke

Finger-Writing Recognition System using Hidden Markov Model
Author(s): Umama Ahmed, and Yishu Zheng

Using Agent-Based Modeling to Understand the Effects of Privately Contracted Security Teams on Pirate Behavior and Maritime vessel Route Selection in the Horn of Africa
Author(s): Rebecca Law

Developing a Simulation for College Students Learning the Rate of Law in Chemistry
Author(s): Yi-Ching Lin

Finger-Writing Recognition System using Hidden Markov Model

Umama Ahmed
Department of Modeling, Simulation and Visualization Engineering
Old Dominion University
uahme001@odu.edu

Yishu Zheng
Department of Modeling, Simulation and Visualization Engineering
Old Dominion University
yzhen003@odu.edu

Abstract - The task of finger gesture recognition is of significant importance due to its wide application in electronic devices. In this paper, a finger gesture recognition system is developed based on Hidden Markov Model to recognize the handwritten letters 'O', 'L' and 'α'. The finger gesture for each of the letters was collected as a form of acceleration data along the three diagonal axis (x, y and z) using the accelerometer sensor of a cell phone. The Matlab toolbox 'HMM' was used to extract features, train the model and identify the finger gestures. It was found that our model was effective for recognizing the letters 'O' and 'α' but not so effective for the letter 'L'. The limitation of the model and several measures to increase the recognition rate of the model are described at the end.

Keywords – Hidden Markov, gesture recognition, transition probability, emission probability.

I. INTRODUCTION

Hidden Markov model (HMM) is a state based modeling technique in which the system is modeled based on the knowledge of the outputs of functions associated with the states of the system. The HMM is built on Markov chains, however, there are some difference in the observables between the HMM and Markov models. In a simple Markov model, the probabilities associated with state transitions are known. Thus, the future states are determined considering the transition probabilities to the future state from the current state. In HMM the current states are unknown; rather the output of functions or sequence of outputs associated with states is known [1].

There are several applications of HMM. HMM is used in pattern recognition for speech, gesture recognition, part of speech tagging, in bioinformatics for predicting and analyzing certain protein and molecular chains. For further discussion of applications of Hidden Markov Models, see [2-4].

In this paper, we will present the results of data analysis of a finger gesture recognition system developed based on the Hidden Markov Model. The team will employ a sample of finger gesture data of different alphabetic letter to build the HMM model, extract the features, train the model, and finally identify each specific gesture using Matlab toolbox of Hidden Markov Model.

The paper will first describe the main objective of the research in section two, followed by the description of data collection in section three. Section four will describe the details about the model, and section five will present the data processing method. Results and analysis will be presented in section six, followed by the conclusion in section seven.

II. OBJECTIVE

The main goal of this project is to build a finger gesture recognition system using HMM toolbox from Matlab [5]. We will attempt to analyze and see how efficient is the model to recognize the finger gestures. The expected outcome of this project will be a gesture recognition system that will be able to identify the specific alphabetic letters defined for this project.

III. DATA COLLECTION

The finger gesture data for this paper contained accelerometer data for specific finger gesture. The accelerometer of a cell phone was employed to collect the finger gesture data (this is an experimental method where a research volunteer was selected to draw the letters on the cell phone; data of each of the letter drawn on the accelerometer was recorded). An accelerometer senses the acceleration data of three orthogonal axes (x, y, z axis) in each gesture for each time stamp. A gesture can be denoted as:

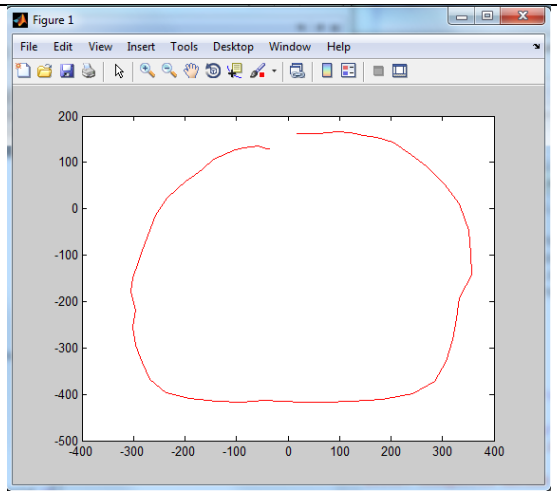
$$G = (ax, ay, az)$$

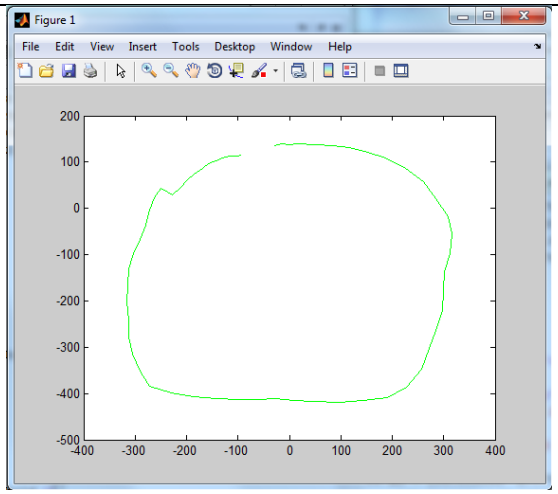
For the finger gesture recognition system, accelerometer data of finger gestures at a specific time stamp are collected for different alphabetic handwritten letter (O, α and L) along x, y and z axis over 60 time

stamps. Therefore, we have 180 data point per each sample data. A total of 10 sample data were collected for each letter; this means, a total of 1800 data per each alphabetic letter.

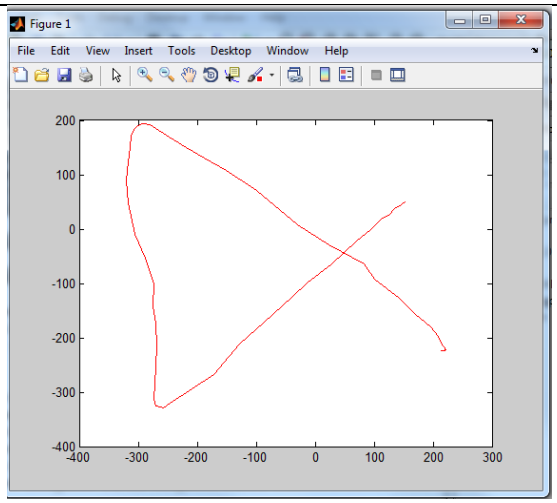
By plugging in the 180 accelerometer data of one data sample, we can obtain the actual finger gesture created by the user. Table 1 displays an example of the data plot for each letter.

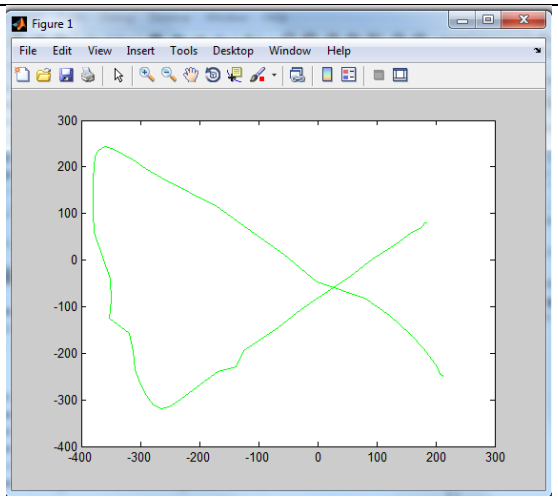
Table 1. Data plot sample



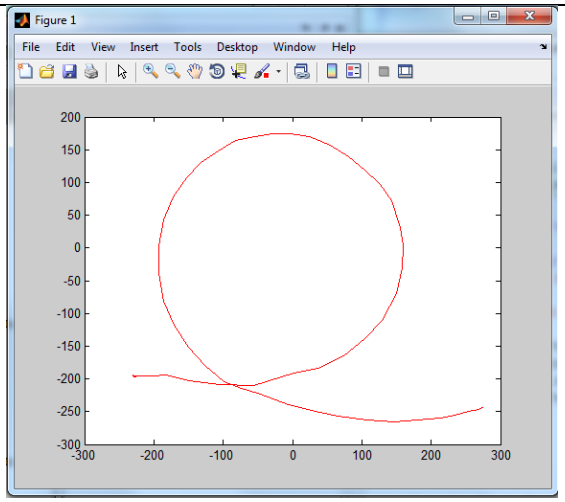


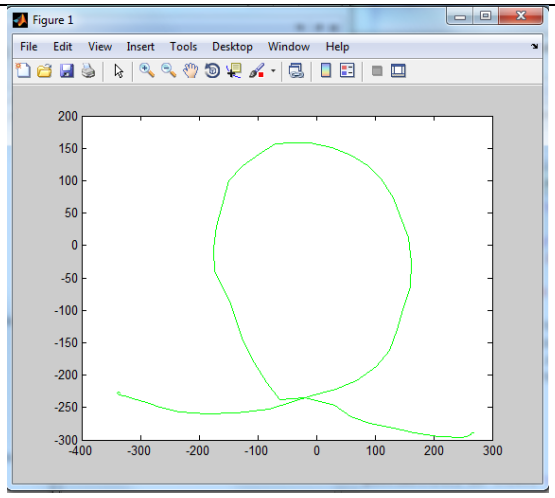
Two sample data points for 'O' hand gesture





Two sample data points for 'X' hand gesture





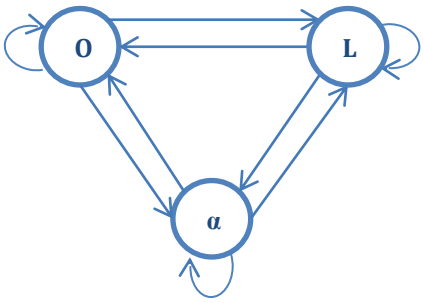
Two sample data points for 'L' hand gesture

We use these data to train our HMM model and to estimate transition and emission probabilities. Finally using another set of sample data, we will test our model to check whether the model can accurately recognize finger gesture.

IV. MODEL DESCRIPTION

Our data set consisted of finger gestures for the three letters 'O', 'L' and 'α'. Thus we have three states from which the sequence of data can come from. The transition matrix is a 3 by 3 matrix. The in depth look to the data observations of the sample data revealed that the data values for these three letters ranged from 1 to 159. Therefore, we have total 159 observations. The emission matrix is a 3 by 159 matrix. The HMM model is constructed as in Figure 1.

Figure 1. HMM model for finger gesture recognition



The States: Q= {O, L, α}

The observations: V= {v_1, v_2,.....,v_{159}}

The next section describes the process of estimating transition and estimation probability matrix.

V. DATA PROCESSING IN MATLAB

V.I. Parameter Estimations

As there are 180 data for values of acceleration along x, y and z for each data sample, we first narrow the data into one value by taking the square root of the sum of their power ($\sqrt{a_x^2 + a_y^2 + a_z^2}$). By doing this, we ensure that each time stamp data has one value, and we obtained a sequence of 10 values per each data sample. We reduced each value into a three unit number, and we round them up since the *hmmtrain()* function of Matlab only can process integer values.

For initial transition probabilities, we assume an even prior value for all three alphabetic letters, so the initial transition matrix is displayed on Table 2.

Table 2. Initial transition matrix

	O	***α***	***L***
O	1/3	1/3	1/3
α	1/3	1/3	1/3
L	1/3	1/3	1/3

The emission probability is calculated by computing the ration of each value in the whole set of values obtained in the trained sample as in Table 3.

Table 3. Emission probability

P(v1|O)	P(v2|O)	…	…	P(v159|O)
P(v1|L)	P(v2|L)	…	…	P(v159|L)
P(v1|α)	P(v2|α)	…	…	P(v159|α)

Using these prior values, we use the Matlab function *hmmtrain()*, which is a Hidden Markov model parameter estimator using prior transition and emission probabilities. The syntax for hmmtrain is as follow:

[ESTTR,ESTEMIT] = hmmtrain(seqs,trans,emis)

This function estimates the transition and emission probabilities for a hidden Markov model using the Baum-Welch algorithm [6]. The parameter *seqs* is a row vector containing different sequences. For our project model, we have 30 data sequences in which X1 to X10 are data samples for 'O' letter, X11 to X20 are data samples for 'X' letter, and X21 to X30 are data samples for 'L' letter. The parameters *trans* and *emis* are the initial estimates for the transition and emission probability matrices. So the result will be a matrix *estTR* for the estimated transition probability, and *estE* for the estimated emission probability [7-8].

After running our model *estimate_paramenters.m*, the estimated transition probabilities are obtained, see Table 4.

Table 4. Estimated transition matrix

	O	***α***	***L***
O	0.8549	0.0846	0.0605
α	0.0377	0.9031	0.0592
L	0	0.055	0.945

The estimated emission probabilities are also generated.

V.II. Test Data

After having the estimated probabilities, we now proceed to test our model in order to verify its accuracy. For this project, we gather a new set of 20 data of alphabetic letters 'O', 'α', and 'L', and we added these data as input in our model to see if it could identify which letter is our set of data representing. We do this by using the function of *dhmm_logprob()* from HMM toolbox to evaluate the log-likelihood of a trained model given test data. From the estimated parameters, the prior transition probability is the one defined as the original matrix *trans;* while the actual transition matrix and observation matrix are the parameters calculated above as *estTR* and *estE*. Our model will then calculate the probabilities:

```
% load a sequence O, α, L
data_letter=[insert data sequence];

% define all the three models
prior_letter=trans;
transmat_letter=estTR;
obsmat_letter=estE;

prior_o= [1];
prior_x = [1];
prior_l=[1];

transmat_o=[1];
transmat_x=[1];
transmat_l=[1];

obsmat_o=obsmat_letter(1,:);
obsmat_x=obsmat_letter(2,:);
obsmat_l=obsmat_letter(3,:);

% find the model that best explains the data
l_O= dhmm_logprob(data_letter, prior_o, transmat_o,
obsmat_o)
% for letter 'O'

l_X = dhmm_logprob(data_letter, prior_x, transmat_x,
obsmat_x)
% for letter 'α'

l_L = dhmm_logprob(data_letter, prior_l,
transmat_l,obsmat_l)
% for letter 'L'
```

From the three answers, the one with the highest values will indicate that our data is most likely to be that letter. We run this model for the 20 values, and the results are displayed in the next section.

VI. RESULTS AND DATA ANALYSIS

The results of running 20 data samples are displayed in Table 5.

Table 5. Results from 20 data sample

	l_O	**l_α**	**l_L**	**Actual letter**	**Estimated letter**	**Assert**
DATA 1	-148.82	-11190.00	-Inf	O	O	1
DATA 2	-149.77	-13260.00	-Inf	α	O	0
DATA 3	-175.21	-18750.00	-Inf	O	O	1
DATA 7	-Inf	-7168.70	-Inf	α	α	1
DATA 6	-Inf	-9487.00	-Inf	L	α	0
DATA 4	-145.99	-14517.00	-Inf	O	O	1
DATA 5	-183.61	-13523.00	-Inf	O	O	1

DATA 8	-Inf	-7130.50	-Inf	α	α	1
DATA 9	-Inf	-3013.10	-Inf	L	X	0
DATA 10	-Inf	-7305.20	-Inf	α	α	1
DATA 11	-Inf	-2433.00	-Inf	O	α	0
DATA 12	-Inf	-Inf	-Inf	L	none	0
DATA 13	-832.26	-15205.00	-Inf	O	O	1
DATA 14	-Inf	-Inf	-Inf	L	none	0
DATA 15	-Inf	-Inf	-Inf	L	none	0
DATA 16	-Inf	-192.24	-Inf	α	α	1
DATA 17	-Inf	-3375.30	-Inf	α	α	1
DATA 18	-Inf	-4019.00	-Inf	L	α	0
DATA 19	-937.93	-12713.00	-Inf	O	O	1
DATA 20	-Inf	-4900.00	-Inf	L	α	0

From the total of 20 samples, 11 letters where correctly found, meaning that the model is 55% effective in guessing the handwriting letters. Taking a closer look, we realized that the model is more effective in recognizing the letters 'O' and 'α' what the letter 'L'; and a specific data analysis showed that 85.71% of the times 'O' is recognized, 83.33% of the times 'α' is recognized, while none of the 'L' letters where recognized, or confused with 'α' letter. The assertiveness of the model is presented in Table 6.

Table 6. Assertiveness of the model

	O	**α**	**L**
Assert	6	5	0
Total	7	6	7
% assert	85.71%	83.33%	0.00%

By looking at the sequences of data set for each letter, we realized that the values letter 'α' are sometimes having similar behaviors of the letter 'L', this can be consider as one of the reasons why 'L' is often confused with 'α' in our model. Also, if we look at the graphical image of both letters, we can see that they present similar patterns in some parts of the letter; therefore, making the recognition more difficult. As a result, we computed a new model for letter recognition, this time only to compare letter 'O' and letter 'L'. We change the prior transition probability *trans* to a 2x2 matrix with even probability 0.5, and we compute the new estimated parameters for *estTR* and *estE*.

Having the new parameters, we take again 20 random data samples and we run the model, the results are displayed in Table 7.

Table 7. Results with new parameters

	l_O	**l_L**	**Actual letter**	**Estimated letter**	**Assert**
DATA 1	-179.90	-Inf	O	O	1
DATA 2	-180.78	-Inf	O	O	1
DATA 3	-190.80	-Inf	O	O	1
DATA 4	-191.50	-Inf	O	O	1
DATA 5	-Inf	-Inf	O	none	0
DATA 6	-Inf	-1451.30	O	L	0
DATA 7	-Inf	-Inf	O	none	0
DATA 8	-Inf	-1143.90	O	L	0
DATA 9	-217.07	-Inf	O	O	1
DATA 10	-182.93	-Inf	O	O	1
DATA 11	-Inf	-5081.30	L	O	0
DATA 12	-Inf	-179.1075	L	L	1
DATA 13	-Inf	-857.2375	L	L	1
DATA 14	-Inf	-175.811	L	L	1

DATA 15	-Inf	-188.2158	L	L	1
DATA 16	-Inf	-175.4708	L	L	1
DATA 17	-Inf	-1033.8	L	L	1
DATA 18	-Inf	-188.5673	L	L	1
DATA 19	-Inf	-Inf	L	none	0
DATA 20	-Inf	-196.6826	L	L	1

From the total of 20 samples, 14 letters where correctly found, meaning that the model is now 70% effective in guessing the handwriting letters. This time, we realized that the model is more effective in recognizing the letters 'L' comparing to the first model. Letter 'L' are recognized 8 of the total 10 samples, and letter 'O' is recognized 6 of the total 10 samples. The assertiveness of the model is presented in Table 8.

Table 8. Assertiveness of the new model

	O	**L**
Assert	6	8
Total	10	10
% assert	60.00%	80.00%

As we can see, this second model increases the effectiveness of the model to recognize the letter 'L'. This is because by eliminating the variable letter 'α', the system can better differentiate the data sequences, especially since the data sequences of 'O' and 'L' present bigger difference in terms of behavior; therefore, making them easier to be recognize. Although, we know that eliminating the variables is not the best way to improve our solution, this was made to test the overall idea to solve this handwriting recognition system.

VII. CONCLUSION

In this paper, we developed a finger gesture recognition system based on Hidden Markov Model. We collected accelerometer data for the finger gesture of the hand written letters 'O', 'L' and 'α' using the accelerometer sensor of a cell phone. Our model was able to recognize the letter 'O' correctly 85.71% of the times. The letter 'α' was correctly recognized 83.33% of the times. However, the model was not efficient to recognize the letter 'L'. The model failed to recognize the letter 'L' and mostly confused it with the letter 'α'. This was due to the fact that the observation sequences of the letter 'L' sometimes have similar behaviors as the observation sequence of 'α'. In order to reduce this confusion, we developed another model with two letters 'O' and 'L', which usually has two distinct observation sequences. This model was able to recognize the letter 'L' 80% of the times and letter 'O' 70% of the times.

Several measures should be taken in order to increase the recognition rate of our model. First, the model should be trained with more training data set to extract more features. Limited number of training data was one weakness of our model. Also the data should be collected over larger time stamps. Although, we already have 60 time stamps, for some similar letters this might not be enough. By increasing the time stamps, it will ensure more feature extraction for each of the gestures; therefore, more accuracy for our model. Finally, the model can be developed using an alternate toolbox which is not limited by any integer restrictions. The integer restrictions in the *hmmtrain()* function in hmm toolbox of Matlab forced us to round up the accelerometer observation data, which might affect the recognition process of the model.

REFERENCES:

[1] McKenzie, R. (2012). MSIM 605 Engineering System Modeling Class Lecture – State Based Modeling, Old Dominion University.

[2] D. H. Rubine, "The automatic recognition of gesture," Ph.D dissertation, Computer Science Department, Carnegie Mellon University, December, 1991.

[3] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, *A Hidden Markov Model-Based Continuous Gesture Recognition System for Hand Motion Trajectory*, International Conference on Pattern Recognition (ICPR), pp. 519-522, 2008.

[4] Schlomer, T., Poppinga, B., Henze, N. and Boll, S. Gesture recognition with a Wii controller, TEI '08 Proceedings of the 2nd international conference on Tangible and embedded interaction Pages 11-14.

[5] Matlab HMM toolbox.

[6] The documentation of Hidden Markov Model in Matlab, retrieved on November 22nd, from http://www.mathworks.com/help/stats/hidden-markov-models-hmm.html.

[7] Mathworks. (2012)., hmmtrain. http://www.mathworks.com/help/stats/hmmtrain.html

[8] Visser, I. (2011). Seven things to remember about Hidden Markov Models: A tutorial on Markovian models for time series. Journal of Mathematical Psychology. Vol. 55 P.403-415

A Conceptual Model Utilizing Evolutionary Game Theory to Explore the Effects of Discount Factors on Interactions Between International Organizations

Rebecca Law, Old Dominion University

In my experience as a student and professional, I often am surprised (and excited) to see theory apply directly real world situation. So often there are striking degrees of variance in real world scenarios that lesson my belief in usefulness of theories. Never was this so apparent until taking an introductory international relations theory course and game theory course. As a sat through each of these courses, I recall trying to always apply one course's material to the other. At times this was difficult at best, but when the concept of shadow of the future was introduced to me in my international relations theory course and mentioned again in my game theory course along with discount factor I was elated to see this link.

The following paper presents a conceptual model utilizing evolutionary game theory to explore the effects of discount factor on interactions between international organizations.

Institutional Theory in International Relations

According to institutional theorists, international institutions matter because they perform four major functions. First, they reduce transaction costs of doing business because meeting through the apparatus of the institution reduces the individual cost for each state. Second, institutions allow for greater transparency and greater verification. For example, states are now privy to the capabilities of others. Finally, institutions increase cooperation by providing a standardization of communication and creating basic norms within the institution.

Shadow of the Future

The concept of the shadow of the future relates to how long the actors believe that they will be interacting. To expand on this premise, presume the actors are in a satiation where they will be interacting for a long time. In this case, the shadow of the future is characterized as "long." Conversely, if actors believe their interaction will be quick and brief, the shadow of the future is characterized as "short." The crux of the shadow of the future is about expectations.

International Relations Institutionalism and the Shadow of the Future

Do institutions lengthen the shadow of the future? Do institutions change how states view one another in their relationship? How?

Institutional theorists contend that institutions do in fact lengthen the shadow of the future. First, most institutions require states to sign publicly. This provides a shared audience of witnesses to the mutual agreement under which states enter institutions. Second, institutions provide an iterative success document, which in turn makes the actors believe

that there is architecture for their new relationship. From this, one can infer that the longer the shadow of the future, the increased likelihood of cooperation. But why? If the shadow of the future is long then states believe that the relationship will be a prolonged relationship. This belief has direct implications for a state's behavior now. A state must carefully consider their behavior and the consequences of cheating now or cooperating now. A further examination of this last statement is necessary.

Let us contemplate a situation within an institutional setting where The shadow of the future is long. States must consider that if they cheat now on an agreement they may be cheated on or punished later. Now, imagine a state in an institutional setting where the shadow of the future is short. States must consider that if they you cooperate now then they may be rewarded. Nevertheless, states can also get away cheating without the prospect of consequences in this scenario due to the brevity of the relationship.

Contrariwise, if a state chooses to cooperate in a short shadow of the future relationship the brevity of that relationship may not warrant time for reward either. The critical point here is that the longer the shadow of the future the more the cooperation now.

Game Theory and Shadow of the Future

Until now the concept of the shadow of the future has been discussed in the context of IR theory, specifically through the lens of an institutional theorist; however, the notion of the shadow of the future transitions seamlessly into the domain of game theory.

The concept of a discount factor- how much a decider prefers rewards now over rewards later- is the key to this transition. The discount factor is represented by the symbol δ, $with\ 0 < \delta < 1$.[1]

Hypothesis

From the knowledge gained on the concept of the shadow of the future, institutional theory and fundamental game theory applications such as δ, the author of this research paper proposes the following hypothesis for further examination:

If states enter an institution under the supposition that the shadow of future will be short and the shadow of the future in reality ends up being long, the effectiveness of that institution to mitigate a problem will be greatly diminished. Conversely, if states enter an institution under the supposition that the shadow of the future will be long and the shadow ends up being short in reality, the effectiveness of that institution to mitigate that problem will not be as diminished as severely as the first scenario.

Conceptual Model Development

The concept for model development is straightforward but requires some imagination on the reader's part. Imagine a problem exists that is global in nature. (Global warming or pollution is one such example.) The magnitude of the problem has become so great that it has now warranted multiple states and institutions enter an institution, or agreement, to address this growing global threat.

For the purpose of this game set up, however, the conceptual model focus is captured at the time when the

agreement or institution is formed. Furthermore, only two fictional institutions are considered.

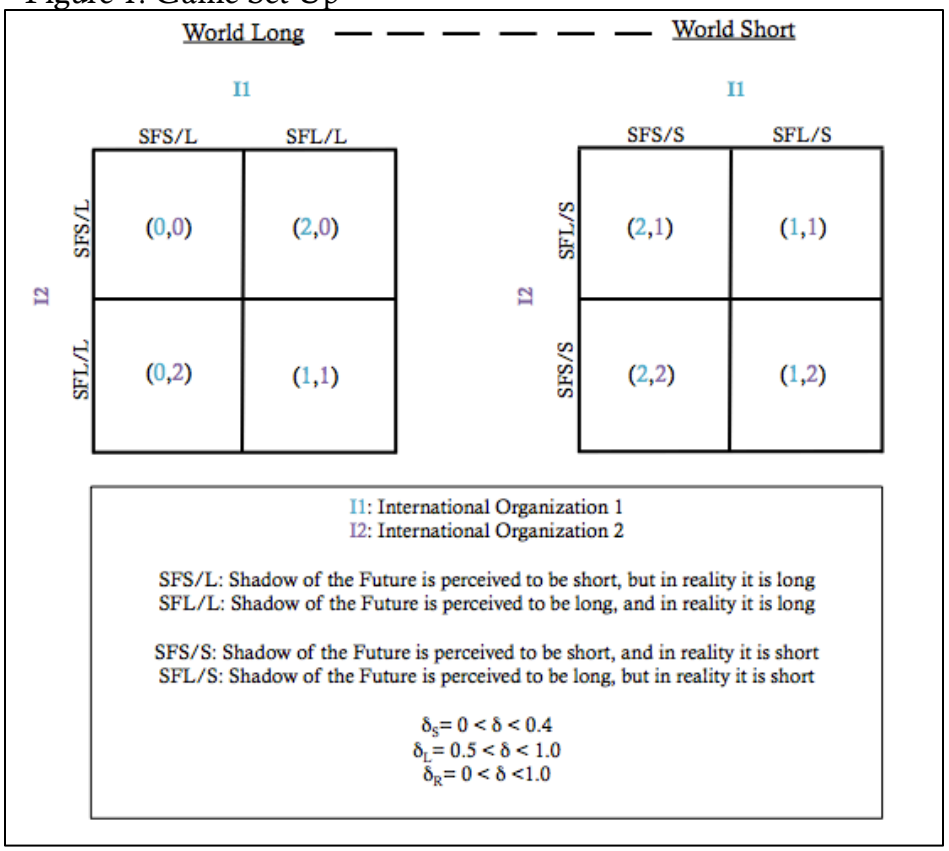
First, two fictional institutions represent the players of the game. The institutions shall be referred to as Institution 1 (I1) and Institution 2 (I2) throughout the rest of the paper.

Second, two distinct states of the world exist, initially determined by nature. World Long is a world in which the nature of the problem the institutions have joined warrants both I1 and I2 to anticipate a long shadow of the future. World Short is a word in which the nature of the problem the institutions have joined warrants both I1 and I2 to presume a short shadow of the future. Within these two states of the world, all possible combinations of I1 and I2's assumptions about the shadow of the future are represented. (see Figure 1.)

In addition, a indefinite Prisoner's Dilemma set up is chosen since interactions between I1 and I2 will be repeated although the state of the world will dictate how many times. A modification of the game to accommodate real world conditions is discussed later in the paper.

Next, various values for δ were assigned to I1 and I2 depending on what their assumption about the shadow of the future is. (see Figure 1.)

Figure 1. Game Set Up



Finally, the payoff structure for each player varies as it depends on which state of the world they exist.

How to Solve the Game

Although the game was not solved at the time this paper was written, further information regarding evolutionary game theory and Grimm Trigger strategy were studied to further refine this model. The following steps should be taken in order to solve this game:

- Calculate the expected payoff of each strategy in each state of the world.
- Determine the probability of each type of player.

- Assume a Grimm Trigger Strategy played by the single player who is incorrect of their knowledge of the state of the world. (i.e player playing SFS/L or SFL/S)
- Calculate the expected utility of each payoff.

Hypothesized Results

Preliminary results are abstract rather than concrete. The author postulates that the closer institutions' values for δ are to each other and δ_R, the greater the ability of these institutions to mitigate a problem. This is perhaps the reason why institutions are better able to increase cooperation. When expectations are similar, the likelihood of cooperation is increased as well as the ability to sustain that cooperation over a long period of time.

Conversely, when the institutions' values for δ are least compatible both with each other and δ_R, the ability of the institutions to mitigate a problem is impacted at its' greatest level.

Implications of Hypothesized Results

The reality of this situation requires an understanding of other factors that may influence the model. For example, a state or institution playing the strategy SFS/L may free ride (off of another state or institution playing SFL/L. In this case, the values associated within each payoff become questionable. For example, an institution that free rides would gain greater utility from free riding because they are not expending as much money and resources on the problem as the other institution; conversely, the institution that is the victim of free riding may in fact gain utility in knowing their intuition is mitigating the problem regardless of the fact that they are doing so at greater expense.

Case Study

The author of this study finds it useful to provide an example of where such research might benefit real world application. As such, a case study drawing upon current events involving maritime piracy off the coast of Somalia and the international community's attempts at mitigating this problem is discussed.

Background

Maritime piracy has existed since the seas were piled for trade. Piracy has traditionally been romanticized extensively by both writers and filmmakers, and has otherwise been consigned to the past. Recently, however, there has been an extraordinary increase in the number of attacks on commercial seafarers and their vessels. Modern-day piracy, specifically in the Horn of Africa (HOA) and Gulf of Aden, has transformed into a business-like industry incentivized by a high reward and low risk. Should this problem remain rampant, the impacts could have disastrous economic, political, and security consequences globally.

Current international counter-piracy measures represent an unprecedented level of international maritime collaboration that has emerged in a relatively short period of time, yet they have singularly failed to dent the incidence and scale of Somali piracy. Several reasons account for this.

First, existing strategies have been compartmentalized by the various agencies involved, which have largely not moved to develop, much less

implement, a uniformed set of goals and objectives.

Second, certain key stakeholders have been conspicuously absent in the formation of comprehensive approaches – notably ship owners who for cost-related and self-interested reasons have unwittingly made their assets highly vulnerable to attack.

Third, there are presently no agreed measures of effectiveness (MOEs) to determine the "success" and cost-utility (or otherwise) of policies designed to enhance maritime security off the HoA.

Fourth, and perhaps most fundamentally, the basic thrust of the response has been premised on a containment strategy that seeks to confront piracy at its end point – at sea – rather than at its root – on land.

It is the belief of this author that states and institutions seeking to eradicate maritime piracy in 2004 underestimated the will and desperation of the Somali pirates. As a result, the international community thought a show of naval force off the coast of Somalia would quickly address this problem. However, today the maritime piracy problem still remains rampant and is a booming economic industry not only off the coast of Somalia but in West Africa as well.

The author believes that states and institutions presumed the shadow of the future for anti-piracy coalitions would be short upon entering the initial agreements and coalitions. This is the reasoning for the overwhelming military response was initiated. However, over time the tactics and resolve of the Somali pirates changed becoming more innovative, sophisticated, and adaptive in their new environment.

Anti-piracy coalitions too adjusted their techniques due to learning. This can be seen by the advancement of attacks from sea to land where pirate camps conduct operations. There is also more consideration for land based approaches to piracy rather than a sea-based approach.

In conclusion, the nature of the Somali pirate problem required anti-piracy coalitions to enter agreements with the understanding that the shadow of the future would be long; however, the author believes this was not the case. As a result, piracy remains as vibrant as ever and anti-piracy coalitions are trying to keep pace.

Limitations

Recognition of limitations in a research project is just as critical as the results of the study. Limitations provide the author with an understanding of how the models can be improved for future research, as well as provide insights into gaps within the existing literature. Limitations to this study include, but are not limited to: time, beginner knowledge of game theoretic models and generalizations.

Time is always a critical dictator in conducting research. The nature and scope of this problem is vast. While the model is imperfect and incomplete, its usefulness is not diminished.

Second, the game theoretic models seem to be pliable. In other words, there was no definitive way to set up and execute the model. The author had a difficult time in choosing what she felt had been the "best" model.

Finally, a number of assumptions and generalizations were made in this study. For example, the game set up only took into account two players, when in reality there are a multitude of players with varying interests, expectations, and utilities that must be considered in order for the model to have greater fidelity.

Future Research and Applications

The author of this study presented the results in an abstract matter. Given a greater amount of time, it would be useful to perform analytical calculations.

However, the usefulness and repurposability of this model to other scenarios presents itself as a strength of this conceptual model and should not be weakened due to the lack of analytical work. This model could be applied to other problems of interest such as budgetary negotiations in the United States Congress for example.

Conclusion

The hypothesized results are preliminary at best, but hopefully have been thought provoking.

As with any research the author conducts, the author tends to walk away with more questions than answers. Questions for consideration include:

- What do the results of this study mean for institutions?
- Are institutions useful? To what extent? Under what conditions?
- Should it be required that intuitions be required to reveal their expectations about the shadow of the future in a signed document upon entering an institution?

It is the belief of the author that a generally accepted assumption about institutions exists. Because institutions increase cooperation, decrease transaction costs and provide some form of standardization and social norms, expectations of institution members are therefore uniform. This assumption is not only fundamentally wrong but also costly. A greater understanding of these concepts is necessary as the world's problems become broader, more complex, and critical.

References

1. Morrow, James D., "Game Theory for Political Scientists," Princeton University Press, 1994.

Developing a simulation for college students learning the rate of law in chemistry

Student: Yi-Ching Lin

Adviser: Dr. Stephen R. Burgin

According to the National Research Center for Career and Technical Education projection STEM careers, researchers projected that the STEM careers will account for 14 % of total occupation by 2018. This is especially important for the disciplines of chemical engineering (6% of engineering major groups) and chemistry (25 % of physical science major groups) which have about a ninety-five percent employment rate (Carnevale, Smith, Stone, Kotamraju, Steuernagel, & Green, 2011). In order to increase the STEM population in chemistry and chemical engineering, many schools have introduced simulation software such CHEMKIN, CHEMSIMUL, HAYS, KINETECUS and MATLAB for training chemistry and chemical engineering students.

Determining the order of reactions and the rate constants are both very important and difficult lessons for students learning chemistry. They are required to integrate many important concepts such as understanding the rate of laws and thermodynamic principals, calculating each compound's mole concentration, and knowing each chemical's property in order to determine the order of reactions and obtain the rate constants. The purpose of this study was to use MATLAB-Simulink to develop a simulation that could be potentially implemented in the future as an aid for college students learning chemistry. In this system, students are required to determine the order of chemical reactions, calculate the rate of constant k, and model initiators and final products' mole concentrations in the entire process through the use of an empirical database.

The results show that this system could be designed in a way that has the potential to help students determine the order of reaction, obtain the rate of constant k, and model initiators and final products' mole concentration. The researchers suggest that using this system might not only help students systemically to integrate their knowledge from both lectures and laboratory courses, but also help students in the control of industrial processes in the future. In addition, the use of modeling and simulation can provide a safe, nonthreatening, and a standardizing learning environment for students to obtain more practical and effective learning experiences (Seoud & Abdallah, 2010).

In conclusion, enhancing students' learning outcomes depends to a certain extent on how they are taught. Computer simulations can be used to support the lab-centered curriculum (Gibbs, & McKeachie, 1990). Using modeling and simulation in conjunction with current lectures and laboratory courses may significantly enhance students' problem solving skills and understanding chemical principles. In the future, researchers are planning to use modeling and simulation in the design and development of chemistry courses more broadly, which may change students' perspective of learning chemistry in the future.

References

Carnevale, Smith, Stone, Kotamraju, Steuernagel, & Green, 2011. Career Clusters: Forecasting demand for high school through college jobs, 2008-2018. Georgetown University Center on Education and the Workforce.

Gibbs, G. & McKeachie, W. J.(1990). *Teaching Tips: Strategies, Research, and Theory for College and University Teachers,* 10th ed.; Houghton-Mifflin: Boston.

Seoud, A. & Abdallah, L., (2010). Two-optimization method to determine the rate constants of a complex chemical reaction using FORTRAN and MATLAB. *American Journal of Applied Sciences 7(4).* 509-517.

Transportation

VMASC Track Chair: Dr. Mike Robinson and Dr. Mecit Cetin
MSVE Track Chair: Dr. ManWo Ng

Simulation Pedestrian and Vehicle Evacuation- A Concept
Author(s): Terra Elzie

Impact of Time of Day on Emergency Vehicle Travel Time Based on GPS Data
Author(s): Khairul Anuar

A Model for Labeling Rank of Each Node in a Directed Transportation Network
Author(s): Abdullah Al Farooq

A Simple and a Fuel Consumption-based Model for Optimal Driving Strategy by Using Probe Vehicle data
Author(s): Ozhan Unal

Simulation Study: Impact of Customs and Check Points on Entity Flow in Seaports
Author(s): Miriam Kotachi

Simultaneous Pedestrian and Vehicle Evacuation - A Concept

Terra L. Elzie, Old Dominion University, VMASC Scholar

Abstract - **The focus of this paper will be on the evacuation of pedestrians and vehicles away from a hazardous area following a no-notice event. The defining issue is addressing the conflicts that will occur between simultaneous evacuation and interactions of pedestrians and vehicles attempting to flee from danger. Hence, the concept of coupling a lane-based evacuation approach for vehicles with the approach of a linear model for the coordination of vehicle and pedestrian flows will be explored. The goal would be to successfully and efficiently evacuate pedestrians and vehicles to safety while avoiding gridlock in the traffic network caused by pedestrian-vehicle interactions. Most related studies analyze either pedestrian evacuations out of buildings or vehicle evacuations ahead of an approaching storm to optimize the clearance time of the traffic network. Yet another type of related study assesses traffic congestion after a pre-planned event such as a football game or a concert at a large stadium. However, only a few studies address the multi-modal evacuation of pedestrians and vehicles. To the author's knowledge, no such papers address multi-modal evacuation due to a no-notice event occurring. This paper explores the concept of how to analyze such a case using modeling and simulation approaches.**

Index Terms - **Concept, Evacuation, Hazardous area, Interactions, Modeling, Multi-modal, Pedestrian Evacuation, Simulation, Simultaneous evacuation, Traffic network, Vehicle evacuation**

I. INTRODUCTION

EVACUATIONS can be associated with a broad range of man-made and natural events. The basic nature of these events, their predictability, frequency, geographic scope, intensity, and other factors define the decisions that must be made by public agencies both in pre-event planning and operational response [1]. According to this ITS document, the seven natural evacuation events are; 1) Earthquake, 2) Flood, 3) Hurricane, 4) Tornado, 5) Tsunami, 6) Volcano and 7) Wildfire. In addition, the six man-made events that could cause evacuation are:

- **Technological** - An event where there is a breakdown in the technological infrastructure such as a power grid failure. Small-scale technological events may include rail-based transit systems requiring emergency evacuation due to a communication or power failure.
- **Hazardous Material** - An event where there is a hazardous material involved and is impacting an area where people are present. This could include an accident on a highway involving a tanker-truck or a derailed train car leaking noxious gas.
- **Nuclear Power Plant** - An event taking place at a nuclear power plant requiring evacuation of the surrounding community.
- **Terrorist Attack** - An unknown event involving the potential harm of people and destruction of property caused by a single individual or coordinated attack by a group of individuals. This may involve a hazardous material (nuclear, biological or chemical) or coincide with a technological event.
- **Dam Break** - An event near or next to a dam (or levy) where the potential for quick and serve flooding of a nearby area is possible.
- **Special Event** - An event such as sporting games, festivals and fairs. Generally these events occur either on a regular basis or there is a fair amount of time in order to plan for such an event.

The natural and man-made types of evacuation events range between planned and no-notice events. Planned events provide ample advanced notice to make a decision on evacuating an area, while no-notice events that occur without warning are generally unpredictable and provide little time to make a decision about evacuating a region. Subsequently, it is easier to make

fully-informed decisions about evacuations for planned events compared to no-notice events since there is a longer lead time to better understand the impact of the event for long-term planning, select an appropriate response and then implement the evacuation plan.

The focus of this paper will be on a no-notice event taking place and the subsequent necessary evacuation of pedestrians and vehicles away from a hazardous area. However, the defining issue is addressing the conflicts that will occur between simultaneous evacuation and interactions of pedestrians and vehicles attempting to flee from danger.

II. PROBLEM

The images of the events that transpired immediately after the collapse of the World Trade Center Towers in New York City following the terrorist attacks on September 11, 2001 are still fresh in everyone's mind even 12 years later. Unfortunately, in addition to the 9/11 terrorists' attacks, others have occurred, including for example, the World Trade Center bombing in 1993 or the 1995 bombing of the Murrah Building in Oklahoma City. From an evacuation standpoint, the common denominator between these emergency situations is the immediate evacuation of all occupants of the targeted and surrounding buildings to flee from the destruction and from the potential for further attacks. It is not only important to have a plan for vehicles to evacuate away from a cordoned off area, but to also establish a plan for pedestrians to evacuate along with those vehicles. However, the interaction between vehicles and pedestrians would cause much conflict by disrupting the flow of traffic, potentially causing gridlock and increasing the probability of vehicle/pedestrian incidents and accidents as pedestrians are intermingled among the vehicles. Therefore, a strategy for evacuating vehicles and pedestrians simultaneously and the successful implementation of a plan is crucial in these emergency situations.

Following the 9/11 attacks, the U. S. Department of Transportation Federal Highway Administration (FHWA) produced a final report on managing pedestrians during the evacuation of metropolitan areas [2]. In this report, the term "pedestrian evacuation" is described as masses of people who leave a suddenly dangerous area in order to reach a safer place and do so on foot. For pedestrian evacuation to be of concern to transportation agencies, it entails the combination of masses of people on foot along with the corresponding congestion of the evacuation of others in private vehicles, always or at times moving along the same routes. Jeffrey Paniati, the Associate Administrator of Operations for FWHA, goes on to say: "Emergencies can occur at any time, any place. We all need to be prepared to take immediate actions to move out of harm's way quickly from wherever we are at the time. The September 11(or 9/11), 2001, attacks on the high-profile workplaces of the World Trade Center (WTC) in New York City and the Pentagon in the Washington, D.C. area, made real the impact of an unexpected, or "no-notice," event in a metropolitan setting. News coverage of the events of 9/11 showed thousands of people leaving the area of the WTC on foot. The evacuation from the borough of Manhattan included not only the typical traffic congestion expected in an evacuation in the United States, but thousands of pedestrians moving along with, or among, the vehicles. However, unexpected emergencies causing people to evacuate an area can result from transportation accidents, hazardous materials releases, earthquakes, flash flooding and other natural and man-made causes."

A study by Cova and Johnson [3] develops a network flow model for lane-based evacuation routing using four separate evacuations that occurred in the downtown Salt Lake City area. The premise of the study was to use lane-based routing to reduce traffic delays at intersections during an evacuation. The scenario that was considered in this paper stemmed from establishing pedestrian evacuation zones (PEZ) and vehicle evacuation zones (VEZ). Cova and Johnson described these zones as follows: "An emergency planning zone containing the incident was blocked to entering traffic by police and designated a pedestrian evacuation zone (PEZ). People in this zone were instructed to proceed out of the zone on foot, abandoning all vehicles. Traffic surrounding this internal zone was instructed to leave the area of its own accord. This represents the surrounding vehicle evacuation zone (VEZ). However during the evacuation, traffic gridlock occurred on the fringe of the PEZ, as drivers were unable to use the sub-network blocked by police to leave the area." Cova and Johnson developed a lane-based evacuation routing approach for the surrounding intersections to alleviate the gridlock surrounding the PEZ and routing vehicles out of this external VEZ. Even though this would require a team of emergency personnel to direct traffic, it holds the potential to reduce intersection delays and network clearing time. However, this study only focuses on the evacuation of the vehicles but does not take into account the pedestrians exiting the pedestrian evacuation zone who now have to enter the vehicle evacuation zone. The developing conflict would have significant effect on the network clearing times at the intersections especially when there are panicked drivers trying to evacuate.

With the above PEZ and VEZ scenario in mind, the 9/11 final report [2] offers three conceptual approaches that share a common objective, to ensure the safety and mobility of pedestrians while minimizing the likelihood that they may contribute to evacuation traffic congestion. Crowd and traffic management techniques can be used to separate vehicle and pedestrian streams, or, taking a different angle, the pedestrian stream can be reduced as the evacuation progresses:

1) Designate and manage separate evacuation corridors for outbound vehicles and for pedestrians.
2) Provide dedicated evacuation transit hubs at the outer perimeter of the evacuation zone to which evacuees can walk.
3) Provide "bus bridges" from where large numbers of people are emerging from the buildings to designated points at the edge of the area being evacuated, where people disembark and begin walking to their destination or find other scheduled mass transit.

In considering each concept above, the first point of managing separate evacuation corridors for outbound vehicles and for pedestrians would be an ideal solution if during the execution pedestrians demonstrate full compliance. It is possible that a percentage of pedestrians will not comply and the issue of pedestrians intermixed among the vehicles would still exist.

The second point of providing dedicated evacuation transit hubs at the outer perimeter of the pedestrian evacuation zone to which evacuees can walk would potentially eliminate the problem of pedestrian evacuating among vehicles. By designating various areas for pick-up, the evacuation of pedestrians among evacuating vehicles would be minimized. However, in emergency situations, the likelihood of individuals congregating and waiting for a bus at one particular location may not be effective even though they are outside of the immediate evacuation zone. Studies on the evaluation of crowd behavior and decision making in these types of emergency situations have been conducted but will not be the main focus of this conceptual paper. Therefore, it would be assumed that a percentage of people will not wait but continue to evacuate by foot beyond the transit hub perpetuating the problem of simultaneous vehicle/pedestrian evacuation.

Finally, the third concept of the 'bus bridge' would address quicker evacuation of people outside of the immediate PEZ once they emerge from their building. However, depending on where the bus requires people to disembark outside of the PEZ would still present the issue of evacuating vehicles and pedestrians interacting.

III. CONCEPTUAL APPROACH

The lane-based evacuation routing strategy presented by Cova and Johnson [3] addresses the transportation problems that arise during an evacuation. By restricting turning options at select intersections, the traffic flow is directed away from the hazardous area while minimizing merging and crossing conflicts among vehicles. The goal of a lane-based routing plan is to temporarily transform an intersection with potentially significant delays into an uninterrupted flow facility. For example, a routing plan might require vehicles in the right lane of an intersection approach to turn right only, while requiring vehicles in the left lane to continue straight, thus minimizing lane changing and merging conflicts. By transforming the intersections within vehicle evacuation zones, the resultant would be fewer traffic delays and lower network evacuation clearing times under moderate to heavy volumes. Therefore, four main objectives were the focus in this study, 1) minimize intersection crossing conflicts, 2) minimize intersection merging conflicts, 3) minimize lane changing along multi-lane arterials and 4) minimize the total evacuee travel distance (shortest network distance). However, the shortest evacuation plan (SEP) criterion introduces a fundamental trade-off between total travel distance and merging; by reducing the number of merging conflicts in an evacuation routing plan, the distance that vehicles must travel to reach an evacuation zone exit increases. However, when vehicles are routed to their closest exit under the shortest network distance assignment, the merging conflicts will be at its greatest. Conversely, in order for merging to be minimized, vehicles must be routed away from their closest exit thus increasing the evacuation travel distance. Nonetheless, as mentioned in the previous section, the evacuating pedestrians emerging from the pedestrian evacuation zone are not considered among the evacuating vehicles.

The results from an applicable study conducted by Zhang and Chang [4] indicate that "neglecting the interactions and conflicts between vehicle and pedestrian flows could significantly over-estimate the evacuation flow rate for both pedestrians and vehicles, especially at links where conflicts exist between the two". Therefore, this study developed an integrated linear model that provides an effective coordination between the vehicle and pedestrian flows during the multi-modal evacuation process. It presents an enhanced model for optimizing the pedestrian and vehicle movements within an evacuation zone that takes into account the conflicts

between congested vehicle and pedestrian flows. By integrating the pedestrian and vehicle networks, the potential conflicts are considered and the optimal routing strategies are generated for guiding evacuees moving toward either their pick-up locations or their parking areas. The model follows two features for guiding and controlling vehicle-pedestrian flows:

1) Realistically represent the networks of vehicle and pedestrian flows and capture their interactions;
2) Integrate the pedestrian network and the vehicle network and compute the optimal departure rate for each intersection approach.

Zhang and Chang give a good approach to modeling and analyzing vehicle and pedestrian interactions, however, the evacuation seems to be more focused on planned special events as they use an illustrative example of evacuating the M&T Stadium in the Baltimore downtown area to test their model. As people emerge from the stadium and walk to their respective vehicles within the numerous stadium parking lots, the dynamic of 'emergency' seems to not be a major consideration although it is alluded to in the paper.

Therefore, exploring the concept of coupling a lane-based evacuation plan for vehicles with the approach of a linear model for the coordination of vehicle and pedestrian flows may be beneficial to obtain an accurate account of a simultaneous vehicle-pedestrian emergency evacuation away from a hazardous area.

IV. CONCLUSION

This paper presents a concept of addressing the need for simultaneous evacuation of pedestrians and vehicles away from a hazardous area due to the occurrence of a no-notice event. By combining a lane-based evacuation plan for vehicles with the approach of a linear model for the coordination of vehicle and pedestrian flows may be a good approach to developing simultaneous vehicle-pedestrian emergency evacuation away from a hazardous area. Most related studies analyze either pedestrian evacuations out of buildings or vehicle evacuations ahead of an approaching storm to optimize the clearance time of the traffic network. Yet another type of related study assesses traffic congestion after a pre-planned event such as a football game or a concert at a large stadium. However, only a few studies address the multi-modal evacuation of pedestrians and vehicles. To the author's knowledge, no such papers address multi-modal evacuation due to a no-notice event occurring. This paper explores the concept of how to analyze such a case using modeling and simulation approaches.

Furthermore, the studies referenced throughout this paper focus on evacuation of large metropolitan areas; however, it would be of interest to explore the need for a simultaneous evacuation plan for a region such as Hampton Roads, VA. This region has a detailed hurricane evacuation plan [5] in place, which an approaching hurricane is considered a 'planned' event given the fact that with advanced technology, there is plenty of notice to set in motion the evacuation plan, if necessary. However, the potential for no-notice emergency evacuations do exist as well. Consider the numerous military bases throughout the region making Hampton Roads a high risk terrorist target. Also consider, with the increase in construction sites in the area, bursting or exploding gas lines would require immediate evacuation of surrounding buildings/area. Even with the existence of the various stadiums in our region, a potential emergency situation may arise where an evacuation of thousands of people is needed. Although, these given examples have low likelihood of happening, the prospect that it could happen makes one contemplate devising a general plan for pedestrian/vehicle evacuation would be beneficial.

REFERENCES

[1] M. Hardy and K. Wunderlich, "Evacuation Management Operations (EMO) Modeling Assessment: Transportation Modeling Inventory", Noblis, Falls Church, VA, ITS Joint Program Office, DTFH61-05-D-00002 (04050002-01), Oct. 2007.

[2] P. A. Bolton, "Managing Pedestrians during Evacuations of Metropolitan Areas – Final Report", Battelle, Seattle, WA, FHWA-HOP-07-066, March 2007.

[3] T. J. Cova and J. P. Johnson, "A Network Flow Model for Lane-Based Evacuation Routing", Pergamon, Salt Lake City, UT, Transportation Research Part A 37 (2003) 579-604, Dec. 2002.

[4] X. Zhang and G. Chang, "Optimal Guidance of Pedestrian-Vehicle Mixed Flows in Urban Evacuation", *Transportation Research Board 90th Annual Meeting*, Washington D.C., Jan. 2011.

[5] "*Virginia Hurricane Evacuation Guide*", Virginia Department of Transportation, Richmond, VA, FEMA, Job No. 12034, May 2012

Impact of Time of Day on Emergency Vehicle Travel Time Based on GPS Data

Khairul (Afi) Anuar
Old Dominion University
Department of Civil Engineering
kanua001@odu.edu

March 24, 2013

Submitted to:
2013 Modeling, Simulation and Visualization Student Capstone Conference

INTRODUCTION

During emergency situations, it is crucial to transport patients to their destinations in a timely manner. To get to the destinations, emergency vehicles are faced with the decisions to determine the path with the shortest travel time. The objective of this paper is to analyze travel time based on GPS data to identify the path with potentially the shortest travel time. As a first step in the analysis, travel times are calculated from emergency vehicles equipped with GPS devices. Since one of the key criteria influencing travel time is the time of day for which the trip occurred, it is important to analyze the extent to which travel time is sensitive (or, alternatively, invariant) to time of day. This analysis is performed by using regression analysis and is used as the basis for determining whether travel times will be fairly consistent across all time periods or will vary across time. Knowing the time characteristics of the path will be important in making route decisions in cases of emergencies.

LITERATURE REVIEW

Research on travel time estimation has been evolving as technologies improve. With increased usage of GPS devices, more data have become available for analysis. Quiroga and Bullock (1998) proposed a methodology for performing travel time studies by combining GPS and GIS data. Taylor et al. (2000) and Shi et al. (2008) extended this GPS-GIS methodology to study traffic characteristics and traffic flow. Li and McDonald (2002) introduced a methodology to estimate link travel time from a single GPS device. To evaluate travel time reliability, Carrion and Levinson (2012) conducted an experiment by placing GPS devices into actual commuter vehicles. A study that is more related to this paper, conducted by Pohorec et al. (2009) looked into the optimization of route selection for emergency vehicles. However, in their study, Pohorec et al. relied on GPS data from regular vehicles. In contrast, the analysis in this paper uses GPS data from emergency vehicles.

METHODOLOGY

Data for this study were provided by the city of Norfolk, Virginia and were collected from April 2012 through June 2012. In Norfolk, all emergency vehicles (EV) are equipped with GPS devices. Through these devices, their positions are monitored by the City's 911 Emergency Control Center. During non-emergency status, their positions are updated for every 500 meters. During emergency status, the position updates are for every 200 meters. For this study, it is assumed that all of the GPS devices in the emergency vehicles were properly calibrated.

The type of EV being studied are the first responders (FR). Fire trucks, another type of EV, are excluded because they do not transport patients. Additionally only FR in emergency status are considered. The path being analyzed is from the east beach area located in northeast Norfolk to Norfolk General Hospital located in southwest Norfolk.

The initial GPS data consists of over five hundred thousand observations. To filter the data, ArcMap was utilized to only include data within the studied travel path. To do so, nodes were created at intersections of major arterials between east beach and Norfolk General Hospital. For each node, ArcMap trimmed the GPS data to only include a search radius of two hundred feet. From these data, several trips were found originating from the east beach area heading to the Norfolk General Hospital. The selection of the links for the regression analysis was based on the routes taken by these FR.

Table 1 summarizes the characteristics of the different links analyzed in this study. Several links were infrequently used by emergency vehicles over the duration of the study time frame. These links are excluded from the analysis because of the small number of observations.

Table 1. Summary of link characteristics

Link	Start Node	End Node	Length (miles)	No. of Lanes	No. of Traffic Signals
1	1	2	0.48	2	3
2	2	3	1.21	2	5
3	3	4	0.72	2	4
4	1	6	2.01	2	3
5	2	5	2.09	2	5
6	3	9	1.73	3	3
7	4	10	0.82	2	5
8	6	7	4.42	1	4
9	5	9	0.91	2	6
10	10	11	1.25	2	4
11	9	12	1.92	2	6
12	7	8	1.89	2	4
13	8	13	0.67	2	3
14	10	14	1.11	1	2
15	14	15	1.15	2	3
16	11	15	1.10	1	2
17	12	16	1.25	2	4
18	13	17	0.74	2	3
19	16	17	0.97	2	7
20	15	18	0.78	1	3
21	18	19	0.16	1	2
22	19	20	0.38	1	2
23	18	21	0.56	1	2
24	19	22	0.49	1	3
25	20	23	0.34	1	3
26	21	22	0.19	2	2
27	22	23	0.31	2	2
28	22	24	1.08	1	5
29	23	25	0.86	2	6
30	17	24	0.59	2	5
31	24	26	0.22	2	2

Note: Highlighted links have insufficient samples for analysis and are excluded from the analysis.

Because travel time is expected to vary by time of day, the FR trips are organized into four categories representing different time periods over a 24-hour interval. These categories are:

AM Peak: 7 am – 9 am
Mid-day: 9 am – 4 pm
PM Peak: 4 pm – 6 pm
Night: 6 pm – 7 am

The location of the nodes and the links are illustrated in Figure 1. The nodes are identified by the numbers inside the circle while the links are identified by the underlined number. The descriptive statistics for travel times on these links are shown in Table 2.

Figure 1. Location of nodes and links

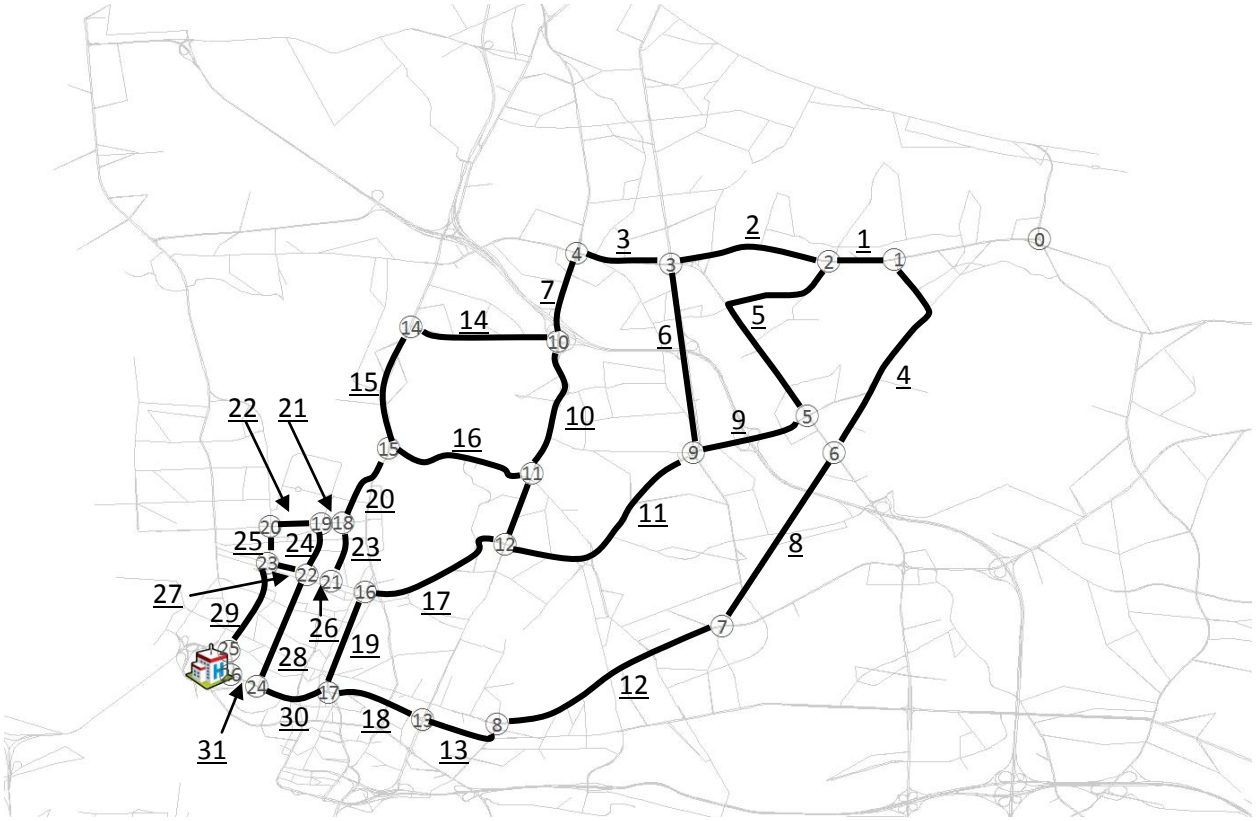


Table 2. Descriptive statistics for travel time (in seconds)

Link	N	Min	Max	Mean	Std. Dev.
1	126	32	195	54.143	24.802
2	85	85	313	131.129	37.892
3	122	54	382	108.681	54.371
4	10	172	325	244.800	59.667
5	6	184	288	221.333	37.521
6	8	146	247	197.875	35.958
7	73	58	195	117.151	32.516
8	2	153	212	182.500	41.719
9	9	69	149	101.222	30.881
10	18	79	167	114.389	23.672
11	31	155	376	244.452	66.834
12	9	162	275	210.444	44.432
13	24	68	250	129.167	55.670
14	225	89	324	159.951	36.846
15	77	77	278	119.065	33.531
16	37	118	774	162.135	105.062
17	55	121	288	177.764	36.493
18	149	65	241	131.577	46.404
19	7	116	156	129.286	15.096
20	76	62	155	96.421	22.336
21	33	17	78	33.242	18.174
22	23	35	147	65.043	30.670
23	13	58	117	83.692	18.979
24	5	69	142	92.000	30.927
25	55	30	167	72.727	29.965
26	163	12	106	29.135	16.026
27	91	20	72	39.088	9.888
28	3	185	275	225.333	45.720
29	244	81	316	170.934	47.246
30	280	66	273	135.886	46.785
31	369	20	124	39.805	18.257

Note: Highlighted links have insufficient samples for analysis and are excluded from the analysis.

Regression analysis is used to determine the extent to which the travel time varied by time of day. Specifically, for each link, travel time is regressed on three dummy variables representing the different time periods of AM peak, mid-day, and PM peak, respectively (with the night period as the reference category). In the regression model, a statistically significant coefficient for a time period would indicate that travel time during that time period is statistically significantly different relative to the night period. The signage of the coefficient would indicate whether the travel time is higher (positive coefficient) or lower (negative coefficient) than that of the night period. Since the night period is expected to have the least congestion and therefore the lowest travel times, the regression coefficients are expected to be positive.

RESULTS and FINDING

Table 3 summarizes the average travel times across each link for the different time periods. As expected, the PM Peak had the highest travel times, followed by mid-day, and night and AM Peak time periods having the lowest travel times. The table also incorporates the results of the regression analysis by indicating the statistical significance of the coefficient for the respective time periods. Using Link 1 as an example, Table 3 shows that travel times during the PM Peak period (of 62.875 seconds) were higher than the night period (48.065 seconds) and that this difference was statistically significant at $p<0.05$. This indicates that FR traveling along Link 1 will experience higher congestion and longer travel times than if they were to travel the link during other time periods. In contrast, travel times on Link 2 averaged 131.129 seconds for all time periods, and regression analysis suggests that travel times do not vary significantly across the different time periods (travel times of 140.500, 125.556, 115.500 and 128.667 for AM Peak, mid-day, PM Peak, and night periods, respectively). This indicates that travel times for Link 2 is invariant and is independent of time of day. Any FR traveling on this link can expect travel times that are reasonably consistent irrespective of time of day.

Table 3 Average travel times by time of day

	Average Travel Time				
Link	**All Times**	**AM Peak**	**Midday**	**PM Peak**	**Night**
1	54.143 (126)	58.300 (10)	54.761 (46)	62.875* (24)	48.065 (46)
2	131.129 (85)	140.500 (4)	125.556 (36)	151.500 (12)	128.667 (33)
3	104.586 (122)	91.167 (6)	115.034 (59)	133.818 (11)	96.804 (46)
7	116.095 (73)	118.000 (3)	113.912 (34)	143.667* (12)	108.375 (24)
11	244.452 (31)	180.500 (2)	242.462 (13)	288.750 (4)	242.500 (12)
14	159.536 (226)	183.692* (13)	158.952 (83)	170.768* (30)	153.510 (100)
15	119.065 (77)	111.500 (4)	125.553 (38)	121.222 (9)	110.000 (26)
16	162.135 (37)	145.333 (3)	184.533 (15)	152.000 (1)	146.833 (18)
17	175.857 (56)	163.250 (4)	176.227 (22)	169.909 (11)	181.526 (19)
18	131.577 (149)	134.800 (10)	128.531 (64)	127.692 (13)	135.016 (62)
20	96.421 (76)	102.400 (5)	93.405 (37)	97.300 (10)	99.458 (24)
21	33.242 (33)	24.500 (2)	42.333 (9)	24.833 (6)	32.375 (16)
22	65.043 (23)	65.000 (1)	74.000 (9)	58.800 (5)	58.880 (8)
25	72.727 (55)	50.000 (3)	78.333 (18)	69.900 (10)	72.542 (24)
26	28.976 (164)	30.316 (19)	28.783 (60)	23.556 (18)	30.334 (67)
27	49.237 (93)	38.286 (7)	38.649 (37)	36.900 (10)	64.410 (39)
29	170.934 (244)	165.294 (17)	181.559* (93)	179.452* (31)	159.709 (103)
30	125.886 (280)	129.000 (22)	137.969 (97)	150.440 (25)	132.838 (136)
31	40.156 (372)	50.581* (31)	39.597 (129)	48.526 (38)	36.885 (174)
Mean	107.653 (2,428)	101.886 (176)	111.969 (950)	115.010 (289)	102.500 (1,013)

Note: N for each time period is shown in parentheses.

* indicates that the mean travel time for that time period is statistically significantly higher than the mean travel time for the

night peak period at $p<0.05$
Unit is second

Based on the regression analysis we can determine travel time variance for the different links. Fourteen links (out of the 19 analyzed in this study) are determined to be time invariant. One link (Link 29) showed higher travel times during the mid-day and PM Peak periods, and one link (Link 14) showed higher travel times during the AM Peak and PM Peak periods. Two links (Links 1 and 7) exhibited higher travel times only during the PM Peak period and another link (Link 31) had higher travel times only during the AM Peak period.

CONCLUSIONS and DISCUSSION

From the regression analysis, emergency vehicles travelling on Little Creek Road (from Azalea Garden to Johnston, Link 1) and on Tidewater Drive (from Little Creek to Thole, Link 7) can expect higher travel time during afternoon rush hour (4:00pm to 6:00pm). Olney Road (from Colonial to Colley, Link 31) exhibited higher travel time is during morning rush hour (7:00am to 9:00am) and on Thole Street (from Tidewater to Granby, Link 14) higher travel times are during the morning and afternoon rush hours. The link with the longest period of higher travel time is Colley Avenue (from 27th to Redgate, Link 29) which is from 9:00am to 6:00pm. The travel time of other links studied in this paper are not affected by time of day.

For the Little Creek Road link, there is no alternative route even though the travel time might be higher during the afternoon rush hor. In this case, first responders have no choice but to expect higher travel time. For city and transportation planners, this is a case where they might consider improving the travel time. This can be done by designating an emergency lane, installing signal preemption (if it is not already installed) or building a hospital near the east beach area.

For the Colley Avenue link, the travel time is higher for a longer time period due to developed areas (higher volume) and with only 2 travel lanes. Again, for the first responders, they have no choice but to expect higher travel time. An alternative route, the Colonial Avenue link did not have sufficient data for a statistical analysis. For city and transportation planners, travel time improvement can be made by replacing street parking with designated emergency lane, increasing number of lanes by making this link a one way direction or installing traffic signal preemption.

For future studies, regression can be performed by including traffic count and traffic signal preemption data. A route choice analysis for an origin-destination would result in a direct comparison of travel times between different routes.

REFERENCE

Li, Yanying and McDonald, Mike. Link Travel Time Estimation Using Single GPS Equipped Probe Vehicle. 5th IEEE International Conference on Intelligent Transportation Systems. Singapore, September, 2002.

Carrion, Carlos and Levinson, David. Valuation of Travel Time Reliability From A GPS-based Experimental Design. Transportation Research Part C. 2012. http://dx.doi.org/10.1016/j.trc.2012.10.010,

Quiroga, Cesar A. and Bullock, Darcy. Travel Time Studies With Global Positioning and Geographic Information Systems: An Integrated Technology. Transportation Research Part C. Volume 6, 1998, pp. 101-127.

Pohorec, Sandi, Verlic, Mateja and Zorman, Milan. Making Applications More Intelligent: A Case Study of Optimizing Route Selection For Emergency Vehicles. 22nd IEEE International Symposium on Computer-Based Medical Systems. Albuquerque, August, 2009.

Taylor, Michael A.P., Woolley, Jeremy E. and Zito, Rocco. Integration of The Global Positioning System and Geographical Information Systems For Traffic Congestion Studies. Transportation Research Part C. Volume 8, 2000, pp. 257-285.

Shi, Wenchuan, Kong, Qing-Jie and Liu, Yuncai. A GPS/GIS Integrated System for Urban Traffic Flow Analysis. 11th IEEE International Conference on Intelligent Transportation Systems. Beijing, October, 2008.

A Model for Labeling Rank of Each Node in a Directed Transportation Network

Old Dominion University

Abdullah Al Farooq
Department of Modeling, Simulation & Visualization Engineering
afaro002@odu.edu

Keywords: Transportation System, Domain Decomposition, graph theory, parallel processing.

Abstract

A real world transportation network contains myriad of nodes(intersections) and links(road segments). Storing this big data in computer memory is too expensive in terms of memory space as there are lots of zeros in node-node adjacency matrix. Sparse matrix representation is a very efficient way to store this big data. To compute different type of operations with such big data, parallel processing reduces the time complexity a lot. In order to partition the whole network in several sub-network, the rank of each node plays a very important role. In this paper a new approach was discussed to get the rank of each node in a directed graph using the data we got from sparse matrix representation. The algorithm discussed here, solved the problem that no matter whether the network is bi-directional or uni-directional, it can compute the rank efficiently.

1. MOTIVATION

Graph traversal for large scale/ real-world data is computationally too expensive [1]. At this point graph partitioning/ domain decomposition plays a very important role to reduce the computational time. Moreover, most of the time the node to node adjacency matrix is too sparse as we consider each intersection of a road to be the source node and there are very few destination points from that source node [3]. So, the ratio of non-zero values and zero values is always too small and near to zero when the network is really big. There have been many researches for storing the matrix in a sparse way. Gary et al. [2] proposed a very efficient way to store the adjacency matrix. In their paper domain decomposition algorithm was employed in such a way that a big network would be divided into different small sub-network and can be computed simultaneously in different processors. In order to decompose the domain, the rank of each node is necessary as the node-node distance matrix is reordered according to the rank for reducing the number of system boundary nodes and for appropriate clustering [2]. Gary et al. [2] considered each link to be bi-directional and thus the node-node adjacency matrix became a bit densed. But for real world large scale data there are lots of uni-directional links which affect the node traversal computational time a lot when the each link in a network is considered to be bi-directional. In this paper a new algorithm was proposed using the sparse matrix representation employed in Gary et al. [2].

2. METHODOLOGY

Gary et al. [1] proposed that instead of building a node-node adjacency matrix one can represent the whole matrix in terms of IA, JA and NZ. For figure 1, the number of outgoing links can be represented in terms of NZ. So NZ and JA will be

$$NZ = \begin{pmatrix} 3 \\ 2 \\ 1 \\ 3 \\ 2 \\ 0 \\ 1 \end{pmatrix}$$

$$JA = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 4 \\ 5 \\ 6 \\ 3 \\ 6 \\ 7 \\ 4 \\ 7 \\ 6 \end{pmatrix}$$

In this paper, a different techniques was used to get the rank of each node and will be stored in the vector RANK which would be calculated from NZ and JA vectors.

RANK(i) = outgoing(i) + incoming(i) ….....(i)

Here, outgoing vector will be as same as the values in NZ vector and incoming vector will be the frequency of a node and that will be summed up with outgoing link of that particular node. The frequency of a specific value can be found from the vector JA. As for example, there are no incoming links in 1, so it will be 0 and there are 3 incoming links in node 6, so incoming links will be 3. so RANK vector will become:

$$RANK = \begin{pmatrix} 3+0 \\ 2+1 \\ 1+2 \\ 3+3 \\ 2+1 \\ 0+3 \\ 1+2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 6 \\ 3 \\ 3 \\ 3 \end{pmatrix}$$

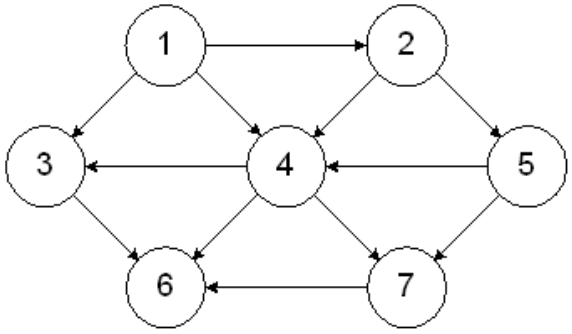


Figure 1: A directed transportation network

3. CONCLUSION

A different type of approach to calculate the rank of each node was proposed here in this paper. As this approach will count the adjacent nodes only when there is a direction toward that node, the renumbering and reordering of nodes should be more appropriate and that will help to decompose the domain more efficiently with less number of system boundary nodes. The more the number of the system boundary nodes, the more computational time would be needed to solve the network.

4. REFERENCES

[1] Duane Merrill, Michael Garland, Andrew Grimshaw. “High Performance and Scalable GPU Graph Traversal” Technical Report CS-2011-05
Department of Computer Science, University of Virginia
Aug, 2011.

[2] Gary Lawson, Shawn Allen, Geoff Rose, Duc Nguyen, ManWo Ng. “Parallel Domain Decomposition Label Correcting Algorithms for real/ large-scale transportation networks on inexpensive laptop with C#/ C++/ MATLAB Computer Environments: Research and Education”. Presented at 92nd Annual Meeting of the Transportation Research Board, Washington, D.C.,
2013.

[3] Shengqi Yang Xifeng Yan Bo Zong Arijit Khan .Towards Effective Partition Management for Large Graphs . SIGMOD’12,May 2012, Scottsdale, AZ, USA.

A Simple and A Fuel Consumption-based Model For Optimal Driving Strategy By Using Probe Vehicle Data

Abstract

With the emergence of the Connected Vehicle technology, i.e., vehicle to infrastructure and infrastructure to vehicle, data from vehicles equipped with this technology provide a better capability to observe traffic flow dynamics, and as a result, enable new applications to improve safety, traffic flow conditions such as coordinating signal timing, and fuel efficiency. The methodology described in this paper uses probe vehicle data and signal timing and generate an optimized vehicle trajectory that prevents the vehicle from stopping at a signalized intersection hence minimizes the vehicle fuel consumption. Methodology in this paper is proposed for a single-lane road leading to a signalized intersection with a fixed signal timing. This study uses a simple model based on equations of motions and a fuel consumption model based on vehicle dynamics data such as instantaneous speed and acceleration levels. The market penetration rate in this study is assumed to be 100%. The method is tested in microscopic traffic simulation software VISSIM.

In this study, there are some assumptions as stated below;

Assumptions:

1- Probe vehicle position is precisely known

2- Uniform arrival rate

3- Constant Speed, constant deceleration rate (V=13.89 m/s, a=3.4 m/s^2 (AASHTO))

4- Reaction time is assumed to be 2.5 seconds

5- Front queue shockwave speed is constant

6- All the vehicles are same in length i.e. 5.0 m and the spacing between two successive cars are taken as 2.0 m.

7- The distance between the first vehicle in the queue and stop line is taken as 1.0 m

8- Cycle length is 90 seconds (45 s is red, 45 s is green)

Methodology

To provide information to probe vehicle to prevent it from Since speed, the queue position and time&space coordinates of the probe vehicle are known, the coordinates of point C can be computed. X-coordinate of point C where the vehicle arrives back of queue can be computed:

- Let the probe vehicle be n^{th} in the queue (assumed to be known). Then the distance between the the front of probe vehicle and the signal is

$X_c = X_{probe} = (n-1) * (5+2) +1$

$X_c = X_{probe} = 7n-6$ .....(1)

where X_c is the space coordinate of point C.

To find the time coordinate of point C, t_c, we use basic distance formula between point D and point C:

$X_{DC} = V_{probe} \times (t_c - t_d)$ .....(2)

where X_{DC} : Dsitance between points D and C

V_{probe} : is the speed of probe vehicle which is 13.89 m/s (50 km/hr)

t_c, t_d : time coordinates of points C and D respectively.

To find the coordinates of point E where the probe vehicle starts moving, we will use shockwave line equation. The shockwave speed is w= -5,3 m/s which is the slope of the shockwave line. Then the equation for this line is

$Y - X_f = w \times (X - t_f)$ .....(3)

where Y and X are space and time coordinates respectively

X_f and t_f are space coordinates of point F where green time starts.

and it is clear that X_f = 1000 m and $t_f = 90 \times i$

where i is the cycle number

Now, since point E satisfies the shockwave speed line equation and $X_E = X_C = X_{probe}$, we can put equation (1) and X_f = 1000 m and $t_f = 90 \times i$ into equation (3) and we get

$(7n-6) - 1000 = -5.3 \times (t_E - 90 \times i)$

$$t_E = \frac{1006-7n}{5.3} + 90i$$ .....(4)

where t_E is the time coordinate of point E.

These calculations will be used in the next section to give information to the probe vehicle by reducing its speed in order to prevent it from stopping and follow the same path after the signal is green.

This can be done in two ways.

1- Each probe vehicle can be informed when they pass at a constant distance i.e. at X=750 m, being the distance 250 m from the signal. In this case, depending on the probe vehicle position in the queue, a speed is calculated and probe vehicle moves with this speed until signal.

2- As an second option, a minimum speed is selected and the distance where the information will be given to the probe vehicle is calculated by using this speed value.

1- Simple Model

In this model, either a desired speed can be calculated or a minimum speed can be determined and the vehicle adjusts its speed to this value. Equations of motions will be used in the calculations.

Calculations

1- Since time coordinate at point G where the information is given to the probe vehicle and the perception reaction time (2,5 sec) are known, time and space coordinates of point A, where the vehicle starts to decelerate, can be computed.

$t_A = t_G + 2.5$

$X_A = X_G + V_{probe} \times 2.5$

where $V_{probe} = 13.89$ m/s

Then

$X_A = X_G + 34.73$ .....(5)

There are two paths between point A and point E. the first one is A-C-E and the latter is A-B-E. Both have the same time difference. Therefore, we can write;

$t_{AC} + t_{CE} = t_1 + t_2$

and let

$T = t_{AC} + t_{CE} = t_1 + t_2$ .....(6)

and from previous calculations, T is known

In order to find t_1 and t_2, distance speed equations will be used

Between points A and B, the probe vehicle is decelerating to desired speed V_{des}. Therefore, we can write

$V_{des} = V_{probe} - a \times t_1$ .....(7)

and the distance between A and B

$X_{AB} = V_{probe} \times t_1 - 1/2 (a \times t_1^2)$ .....(8)

Between points B and E, the probe vehicle is moving with constant speed V_{des}. Therefore, we can write

$X_{BE} = V_{des} \times t_2$ .....(9)

Putting (7) into (9), we get

$X_{BE} = (V_{probe} - a \times t_1) \times t_2$ .....(10)

and from the figure

$X_{AB} + X_{BE} = 1000 - X_A$ .....(11)

putting (8) and (10) into (11), we get

$V_{probe} \times t_1 - 1/2 (a \times t_1^2) + (V_{probe} - a \times t_1) \times t_2 = 1000 - X_A - X_E$ .....(12)

By putting (6) into (12), we get

$V_{probe} \times t_1 - 1/2 (a \times t_1^2) + (V_{probe} - a \times t_1) \times (T - t_1) = 1000 - X_A - X_E$ .....(13)

By solving (13), t_1 and t_2 can be found. Then putting t_1 into (7) V_{des} is computed for the probe vehicle.

2- If a minimum speed is selected then in the above equations V_{des} is known. By using the sama equations, point G where the information is provided can be found for each vehicle.

2- Fuel Consumption Model

An objective function will be defined in order to minimize fuel consumption. This function will compute the fuel consumption of the vehicle in deceleration mode (until it has a constant speed), in constant speed mode and in acceleration mode (until it has its initial speed) as soon as it gets the information.

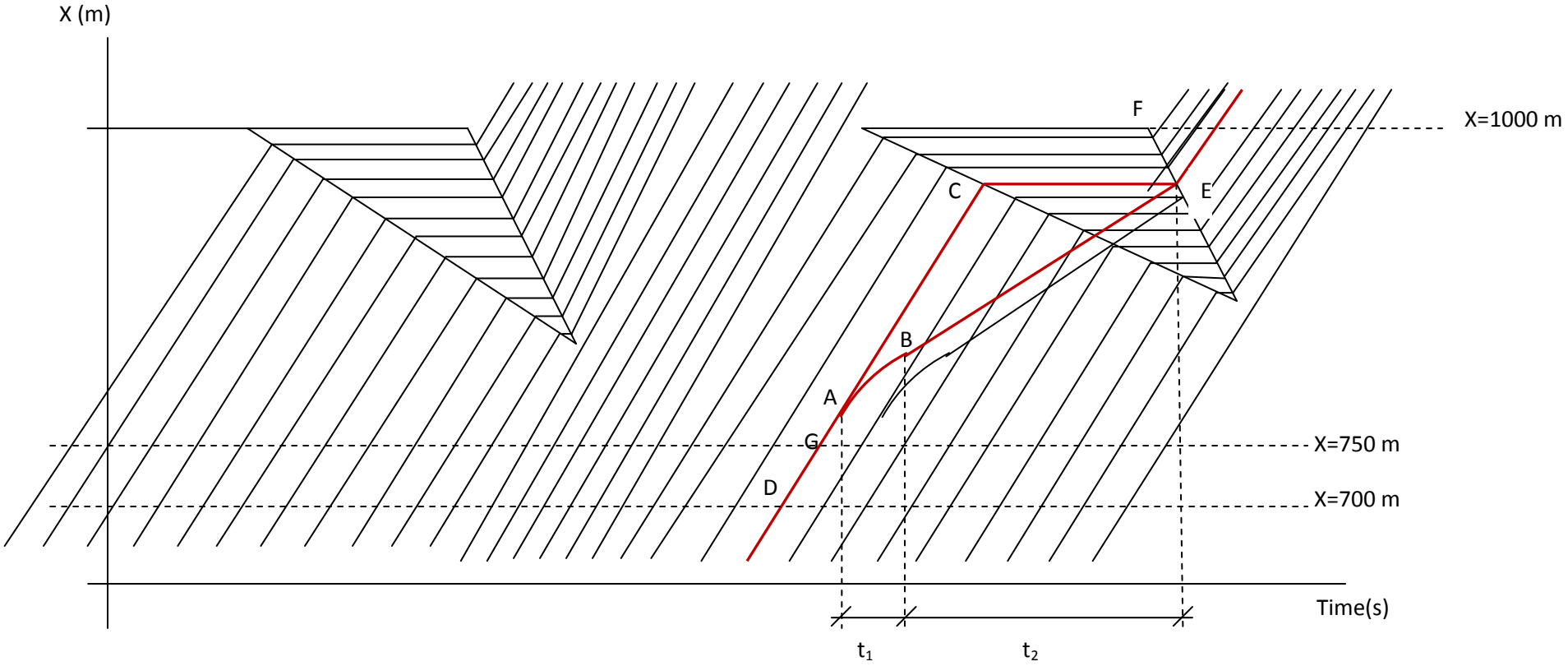


Figure : Vehicle Trajectory At A Signalized Intersection

Simulation Study: Impact of Customs and Check Points on Entity Flow in Seaports

Mariam Kotachi
Department of Engineering Management
Old Dominion University, Norfolk, VA

Gaith Rabadi, PhD
Department of Engineering Management
Old Dominion University, Norfolk, VA

I. Introduction

The main intermodal resources in the world since a long time are ports, where all the other kinds of transportation meet (ships, trucks and rail) to exchange cargo. The United States alone has 361 seaports which are the gateways for more than 80% of the foreign trade; the United States is the world's largest importer and exporter [1].

The basic process in the ports starts by the arrival of ships trucks and rails. All cargo and containers get inspected upon arrival. Containers are then removed by cranes or other methods and sometimes exchanged by loaded or empty ones, loaded containers are either moved to storage area by straddle carriers or moved from one transportation method to another [2] [3].

Security measures and customs delays will significantly influence the port's operations and might lead to spoiling schedules, longer queues and delays in flow of materials throughout the port. Government agencies including customs administrations and control have faced many challenges regarding managing the increasing growth trade, because they have to guarantee law compliances and facilitate lawful trades in the same time. In such cases, it is very hard to keep track of gatekeeping; it is not possible to manage the law compliances of 100 percent control of all transactions, because of all the fast growth of the international trade value and volume plus the limited resources of customs administrations. And because of that some of the new customs control systems follow selectivity approach and risk management [4].

Many researches have been conducted to perform safety measures on ports, by creating simulating models to anticipate the current situation in order to prevent delays and reduce cost. A study has been made about schedule loading operations in container terminals; the method integrates optimization algorithm and evaluation function of simulating model, the main events are: initializing container sequence according to some dispatching rule, then the sequence will be improved by genetic algorithm, and the objective function of a given scheduling scheme will be evaluated using simulation model. At the same time another model is designed to foresee objective function and to remove any probable poor solutions in order to decrease the simulations model running time. Many tests have proven that, scheduling problem of container terminals can be solved by simulation optimization methods [3].

Another popular port related study is the berth allocation problem, which is the assignment of quay space and service time to containers that need to be unloaded or loaded at the terminal. A study has been made to analyze the berth allocation under uncertain arrival time or operation time. It studies the proactive approach to develop an initial schedule that incorporates a degree of expectation of uncertainty during the schedules executions and also studies the reactive recovery strategy which adjusts the initial schedule to handle accurate situations with the lowest penalty cost of deviating from the initial schedule [5].

In the last couple of years it has been proven that modeling and simulation is the best method to help design an intact system that can be considered and looked at as a whole. Modeling and simulation is the best method in problem solving and investigating supply chain nodes, and has been used in many research papers to manage and design ports, terminals and containers [6].

This paper addresses any port in general, but mainly will focus on ports with some customs rules and intermodal node port in the transportation of network, where cargo changes modes of transportation from a ship to an inland transport mode and vice versa.

Data will be collected and provided by the port's management system, for the simulation model development and the result analysis, but will not be reported in this paper due to the confidential nature of such information.

II. Methods

In this work a discrete event simulation model will be developed using Arena to model the movement of incoming/outgoing ships, containers, trucks and rails in a port and to evaluate the impact of customs, security checkpoints and inspections on the flow of the port. Also designing and developing the best approach to prevent any interruptions and expensive consequences. Simulated entities will include: ships, containers, trucks, rail, personnel, cranes and containers carriers. There will be flow of ships, trucks, rails, and the flow of loaded and empty containers among all three of different transportations modes. The resources in the model will be: cranes, straddle carriers, storage and security check points. The model will include the following processes: loading containers, unloading containers, getting containers coming in from the inland and from the sea check in the security check points and moving containers to storage areas.

The proposed simulation will study different scenarios regarding the customs delays or security measures to understand their impact on the port progression. Example of some situations can be some of the following:

- Container storage spaces and policies
- Containers traffic movement throughout the port
- Queue lines for ships, trucks, rails and carriers
- Scheduling problems
- Berth allocation

III. Conclusions

This paper is mainly concerned with designing a discrete event simulation model to find the impact of security check points and customs delays on the flow of entities in the port. Required data is to be collected and statistically analyzed to provide input distributions for arrival and service times as well as parameters for the different stations. In addition, output analysis will be conducted on the results and sensitivity analysis will be used to determine the impact of each parameter on the overall flow and to provide a platform for what-if-scenarios.

References

[1] T. G. Martagan, B. Eksioglu, S. D. Eksioglu and A. G. Greenwood, "A SIMULATION MODEL OF PORT OPERATIONS DURING CRISIS CONDITIONS," *Proceedings of the 2009 Winter Simulation Conference*, pp. 2832- 2843 , 2009.

[2] G. Rabadi, C. A. Pinto, W. Talley and J.-P. Arnaout, "Port recovery from security incidents: a simulation approach," *Bichou, K, Bell, MGH and Evans,* vol. Chapter 5, pp. 83-94, 2007.

[3] Q. Zeng and Z. Yang, "Integrating simulation and optimization to schedule loading operations in container terminals," *Computers & Operations Research,* vol. 36, pp. 1935-1944, 2009.

[4] J. Biljana and A. Trajkova, "Risk management and Customs performance improvements:The case of the Republic of Macedonia," *Social and Behavioral Sciences,* vol. 44 , p. 301 – 313, 2012 .

[5] L. Zhen, L. H. Lee and E. P. Chew, "A decision model for berth allocation under uncertainty," *European Journal of Operational Research,* vol. 212, no. 1, p. 54–68, 2011.

[6] X.-l. Han , Z.-q. Lu and L.-f. Xi, "A proactive approach for simultaneous berth and quay crane scheduling problem with stochastic arrival and handling time," *European Journal of Operational Research,* vol. 207, no. 3, pp. 1327-1340, 2010.

Business and Industry

VMASC Track Chair: Dr. Andy Collins
MSVE Track Chair: Dr. Jim Leathrum

Exploring the Geospatial effect on Foreclosure Contagion Using Agent-Based Modeling and Simulation
Author(s): Daniele Vernon-Bido, Andrew Collins, and Michael Seiler

Popularity or Proclivity? Understanding the Impacts of Instrumental and Intrinsic Preferential Attachment on Social Network Formation Employing Agent-Based Modeling
Author(s): Xiaotian Wang

Investigating Theoretical Frontiers in International Relations
Author(s): Callie Davis

The Induction of Community Dynamics in Online Social Networks
Author(s): David Wright

A Conceptual Model Utilizing Evolutionary Game Theory to Explore the Effects of Discount Factors on Interaction Between International Organizations
Author(s): Rebecca Law

Reducing Stakeholder Fatigue: Integrating Governing Bodies, Committees and Working Groups Contribution to Risk Management
Author(s): David Flanagan, Andrew Collins, and Barry Ezell

Determinants of Seafarer Fatal and Non-fatal Injuries in Container Vessel Accidents
Author(s): Yishu Zheng

A Study of Contagion Spread Among a Finite Human Population on a Naval Vessel
Author(s): Elizabeth Rasnick, Jimmy Hilton, Frederick Guy Wilson

EXPLORING THE GEOSPATIAL EFFECT ON FORECLOSURE CONTAGION USING AGENT-BASED MODELING AND SIMULATION

Daniele Vernon-Bido, Andrew Collins and Michael Seiler

***Abstract* – It is difficult to define the spillover effects of foreclosures on neighboring property values. Some of the variables that affect the neighbors are unsightly neglect, increased crime and below market sales. However, the effect is not the same from neighborhood to neighborhood. Some of this variance is due to the spatial layout and density of the area. This study uses agent-based modeling and simulation (ABMS) to explore the foreclosure contagion differences in a geospatial context. Emergent behaviors are revealed when the density of neighbors in modified.**

I. INTRODUCTION

There is an old adage that recounts the three most important aspects of real estate are location, location and location. Real-world properties have the unique attribute that no two properties can be exactly the same because they cannot share the same physical location. Wilhelmsson notes that "spatial econometrics explicitly accounts for the influence of space in real estate, urban and regional models [11]." It is this spatial concept that has led to the study of the foreclosure contagion effect. The foreclosure contagion effect is defined as the reduction in property values due to neighboring foreclosures [1].

Foreclosed properties sell at lower prices than nearby non-distressed properties [7] and estimating the actual discount of foreclosed homes is subject to debate. However, the sale price of the foreclosed home is taken into account when nearby appraisals are made. Agent-based simulation is a non-traditional way to study of the effects that the foreclosure discount can have on neighboring properties. While this approach was been previous studied by Gangel, Seiler and Collins [1], it utilized a grid system with a homogeneous density of neighbors to calculate the effect. This study will attempt to recreate the agent based model utilizing geographical shape files to represent the physical layout of an existing area. The goal is to understand differences neighborhood density plays in the foreclosure contagion and the community designs that are most influenced by the contagion effect.

II. BACKGROUND/LITERATURE REVIEW

The hedonic regression model is a method that allows for heterogeneous units to be separated into attributes [4] and thus has been appropriate for developing real estate pricing models. The attributes are individually evaluated allowing for indices to be created for aggregate comparisons. Housing prices also demonstrate spatial autocorrelation which, according to Osland diminishes with increasing distance [10].

Likewise, a number of studies have concluded that distressed and foreclosed single family homes have an effect on the value of nearby homes. Immergluck and Smith explored this foreclosure contagion using a hedonic regression model. They reviewed the effect of foreclosures on homes in radius of $1/8^{th}$ of a mile and $1/4^{th}$ of a mile to understand the role of proximity in the housing value change [6]. Immergluck and Smith concluded that on an average sale price of $164,599, the effect of foreclosures within $1/8^{th}$ of a mile was a decrease of 1.136%; between $1/8^{th}$ and $1/4^{th}$ of a mile the decrease reduced to 0.325% indicating that proximity to a foreclosed property is a factor. However, a major assumption in this method is that the housing density in each segment is the same [6].

Other studies, such as Harding, Rosenblatt and Yao, estimate the contagion effect to be considerably smaller than Immergluck and Smith contend [5]. Harding et al. estimate a maximum effect of one percent. However, they agree that distance is still a strong factor. Harding et al. maintain that the strongest contagion effect is with 300 feet of the foreclosed property and that properties $1/8^{th}$ mile away would have a maximum effect of 0.5% [5]. Kobie and Lee modify the methodology to utilize a spatial hedonic model [8]. They studied the time and

space effect of foreclosures in Cuyahoga County, Ohio and introduced the concept of a "face block" as the measured significant distance in contrast to an arbitrarily set distance and a straight line buffer. The "face block" consists of all the houses on both sides of the street from intersection to intersection. [8] Harding et al. and Kobie and Lee note that a lack of maintenance is one cause of devaluation in properties near a foreclosed property. [5, 8] "In addition to normal depreciation, many properties undergoing foreclosure experience gross neglect, abandonment and vandalism which significantly alter their exterior appearance [5]." Kobie and Lee's face block limits devaluation to those homes in a visual proximity [8].

Gangel et al. continue to explore the effects of distance and time on property values in the same community as foreclosed properties [1]. Their study takes a bottoms-up approach using Agent-Based Modeling and Simulation (ABMS). The focus of Gangel et al. is not to measure the effect of foreclosure contagion but to determine the impact that the range of values will have on varying market conditions over time.

This model expands the Gangel model to include the physical layout of neighborhoods. The model attempts to explain one possible cause for the expansive differences in the measured contagion effect. The next sections will describe the model design, the implementation and the results of the model and then present a conclusion.

III. MODEL DESIGN

A GIS (Geographical Information System) real estate foreclosure model was designed to study the variance that a spatial layout has on the foreclosure contagion effect. The agents in the GIS structured real estate model are the various real estate properties; thus the behaviors that the system will act upon are the behaviors of the properties. The key feature of this model design is the use of the shape file which determines the spatial layout of the agents. This model uses a property layout shape file from the northern section of Virginia Beach, VA. It has a perimeter of approximately 11.7 km or 7.24 miles and consists of 1,777 properties.

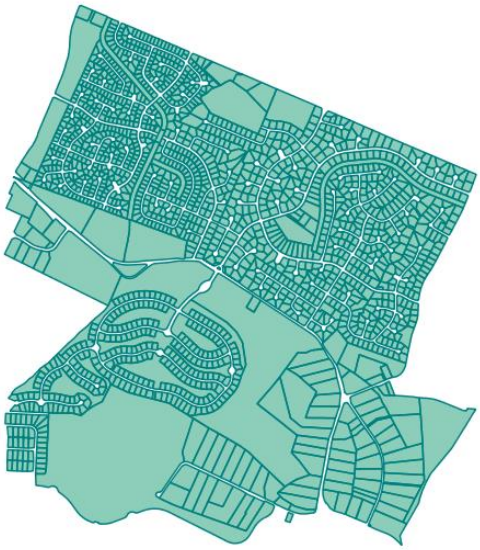


Figure 1: Property layout for northern section of Virginia Beach

The properties have several key functions that govern their behavior. First, instrumental to all agent based modeling system, is the concept of neighbors. Each agent must determine who its neighbors are within the system. A defining feature of real estate properties as agents is that the agents remain stationary therefore the neighbors do not vary. The property valuation is a critical feature for each property. The appraisal process reviews sales of neighboring properties to determine the current value of a property. Property sales are a vital component of the model. However, the dynamic of property sales and listings are not the focus. The model creates a simplified version of the buying and selling process as it is an important part of pricing and foreclosure determination. The last major component is the foreclosure function. There are a number of other components but the following section will provide greater detail of these four primary functions.

A. Neighbor Identification

The neighbor identification process utilizes the geographic coordinates of the real estate property to find all properties within a specified radius. The centroid point is then determined for each property and a bird's eye distance between the property and each of its neighbors is calculated. Utilizing this function eliminates the assumption of uniform

density that exists in the Immergluck and Smith model [6].

B. Property Valuation

Though most real estate models take a hedonic approach to the appraisal of a property, this model for simplicity assumes each dwelling to be homogeneous. The appraisal function generates the current value of each property every month. From the Gangel model [1], the Appraised Value (*AV*) is then designated as the weighted sum of the sales price (p_i) of neighboring properties using the following formula

$$AV = \sum_{i=1}^{n} \frac{p_i}{2}\left(\frac{\Delta d_i}{\sum_{i=1}^{n}\Delta d_i} + \frac{\Delta t_i}{\sum_{i=1}^{n}\Delta t_i}\right) \quad (1)$$

where Δd_i and Δt_i are the differences between the maximum distance constraint (d_{max}) and the actual distance between the properties and the maximum time constraint (t_{max}) and the actual time since the sale of the property respectively.

$$\Delta d_i = d_{max} - d_i \quad (2)$$

$$\Delta t_i = t_{max} - t_i \quad (3)$$

However, if there were no neighboring properties sold within the maximum time constraint, the property is valued at the prior month average price of the entire market. After an appraised value is calculated for a property, the model determines if the value should be discounted based on foreclosures within the specified radius (d_{max}).

C. Foreclosure Probability

The foreclosure effect is a stochastic function that executes monthly and takes into account the equity ratio, the current monthly payment, the investment perspective and the possibility of a catastrophic event. The equity ratio (ER) is calculated as the appraisal value (AV) divided by the outstanding loan balance (LB).

$$ER = \frac{AV}{LB} \quad (4)$$

Equity ratios greater than or equal to one have no effect on the probability of foreclosure, $P(F_{ER})$, since the property value meets or exceeds the debt obligation.

$$P(F_{ER}) = 0 \quad (5)$$

Properties with negative equity or an equity ratio below one are considered to have an increase in the foreclosure probability, $P(F_{ER})$, designated by the following equation:

$$P(F_{ER}) = (1 - ER) * C_{ER} \quad (6)$$

C_{ER} represents the scalar constant effect for the equity ratio. The value of C_{ER} will range between 0 and 1.

Buyers purchase a home with an anticipated monthly payment. If the current monthly payment exceeds the original fixed monthly payment, as can happen with an adjustable rate mortgage, the new payment will be a factor in the foreclosure probability. The payment ratio (PR) is the current monthly payment (CMP) divided by the original fixed monthly payment (FMP)

$$PR = \frac{CMP}{FMP} \quad (7)$$

thus the payment ratio foreclosure probability, $P(F_{PR})$, is given as

$$P(F_{FR}) = (1 - FR) * C_{FR} \quad (8)$$

C_{PR} represents the scalar constant effect for the payment ratio. The value of C_{PR} will range between 0 and 1.

Renter-occupied properties carry an additional probability of foreclosure if the rent does not cover the mortgage payment. For simplicity, this model assumes that a high growth market the rent will be below the mortgage payment increasing the likelihood of foreclosure and in a low growth market the rent will exceed the mortgage payment reducing the likelihood of foreclosure. Owner-occupied homes will not have an effect on this probability factor. Since this is only a factor in the foreclosure probability, it is designated as a positive or negative scalar value, C_{IR}, or zero in the case of owner-occupied properties.

$$P(F_{IR}) = \begin{cases} High\ Growth\ Market = C_{IR} \\ Low\ Growth\ Market = -C_{IR} \\ Owner - Occupied = 0 \end{cases} \quad (9)$$

The final element to the foreclosure probability is the probability of a catastrophic event. A catastrophic event is defined as an external event that results in the inability to maintain payments. Examples of catastrophic events are job loss, death, separation and divorce. The probability of a catastrophic event is represented as a constant $P(C_{CE})$ and is applicable to all properties.

The probability that a property will go into foreclosure is the sum of all probabilities.

$$P(Foreclosure)=P(F_{ER})+P(F_{PR})+P(F_{IR})+P(F_{CE}) \quad (10)$$

D. Sales Determination

Properties that have a positive equity ratio are listed with a higher probability as Genesove and Mayer [2] determined that people are less likely to list properties that are underwater. All properties have the potential to list; however, the higher equity ratio makes the property more likely to list. Following the Gangel model, the property sale function is an aggregated listing and selling function that executes monthly. Each property is given a listing probability based on its equity ratio. The expected number of listings is determined based on the market conditions – high growth or low-growth. The listing impact is a representation of the supply and demand of the local market. In this model, the expected listings (*EL*) are determined by the number of neighbor count (*NC*) and the market-based listing factor (*MF*). The observed listings (*OL*) are a count of the number of neighboring properties that are listed during the month and the listing impact (*LI*) is a linear function of the expected listings, the observed listings and a scalar constant (C_{LI}). The listing impact is then factored into the final selling price.

$$EL = NC * MF \quad (11)$$

$$LI = 1 + (EL - OL) * C_{LI} \quad (12)$$

IV. MODEL IMPLEMENTATION

Repast Simphony is a Java-based ABMS platform. Its integrated use of GeoTools software allowed for the use of shape files as agents while the ABMS framework managed the scheduling functions. The shape file used is the northern section of Virginia Beach. The area is used to represent any suburban type environment with varying subdivision layouts. The purpose is to determine the effect of a heterogeneous density within a given radius on the foreclosure contagion effect. Shape files are immutable. Therefore the model reads the shape file then writes a temporary file which includes all of the GIS details then adds the property and loan details. The initial conditions are the same for all properties. The functions execute in discrete time steps equal to one month. In each step the property value and loan information is updated as well as listed, sold and foreclosed information. The figure below (Figure 2) shows a magnified view of the shape file after a complete simulation run. Outputs are created for each time step.

Figure 2: Expanded view of shape file neighbor depicting foreclosed properties after simulation run

The model is implemented in a single run of 1000 time steps to create baselines for comparison. It is then run in batch mode varying the distance component 10 times.

V. RESULTS

Previous findings concur that there is a foreclosure contagion effect. However, there is significant disagreement on how wide spread the effect is. This model reviews the foreclosure contagion effect as a function of the size of the neighborhood used to determine the effect. The model uses the physical layout of the real property to observe if there are any emergent behaviors. Varying radii are applied to the simulation runs to alter the number of neighbors and compares the effect on the property value.

The foreclosure discount is -0.0657 and a foreclosure sale time of 7 months for the model simulation runs. The metric used to compare these runs is the property value. A single simulation run is completed for each boundary condition. That is, the simulation is run using a radius of 1m. This reduces the number of neighbors to zero and uses the global average to determine property values. In this scenario, the property values increase exponential, shown in Figure 3, since the foreclosure discount has not effect on any neighboring properties.

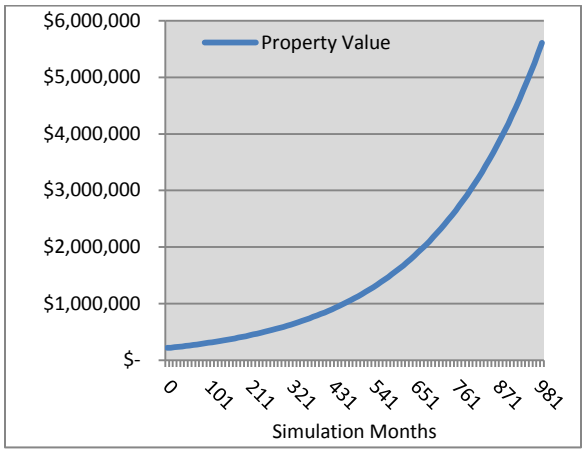


Figure 3: Average property value without including neighbors

Next the simulation is run with all neighbors having a weighted effect on the property value. This is equivalent to every foreclosure, regardless of proximity, affecting the property value of all other homes. This is accomplished by setting the radius to 10,000m. The market in this scenario quickly crashes as properties lose all value (Figure 4).

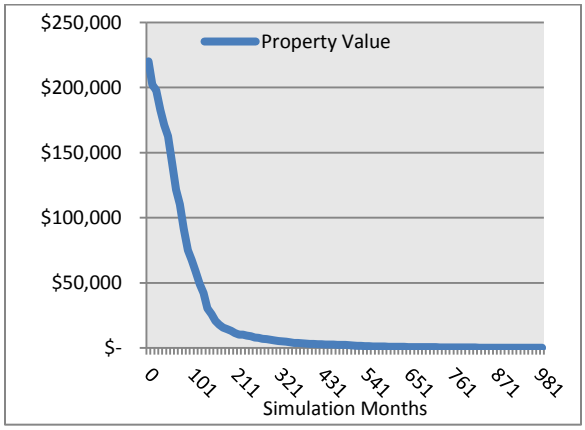


Figure 4: Average property value weighting all neighbors

Finally a baseline run using a radius of 250m which is roughly equal to one city block is executed. The average number of neighbors for this is 101.

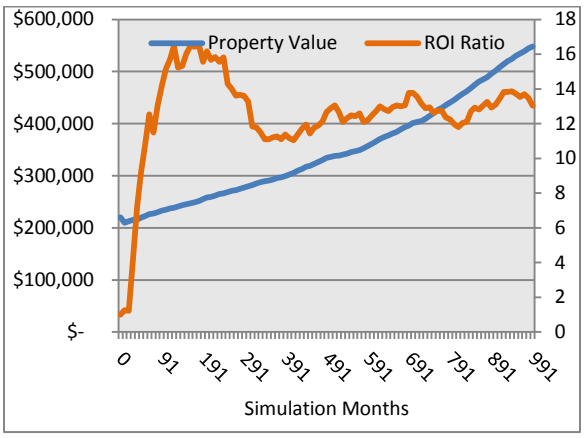


Figure 5: Average Property Value and ROI Ratio for Baseline Simulation Run

The property value as shown in Figure 5 displays a steady increase with minor adjustments through the life of the simulation. The ROI rate, however, varies as buying and selling activities occur. The percentage of foreclosures varies from less than 0.2% to above 1% over the life of the simulation as shown in Figure 6.

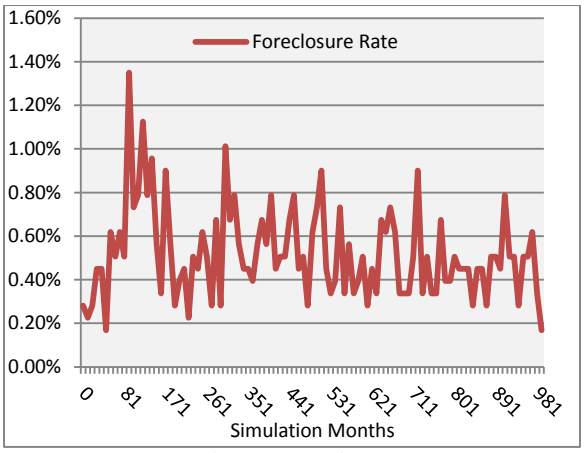


Figure 6: Percentage of Foreclosures for Baseline Simulation Run

Examining the model without any neighboring foreclosures impacting property values and with the entire system impacting property values, the extremes of uncontrolled growth and market demise are observed. Since the market has not assumed either of these conditions, it can be surmised that the foreclosure contagion effect exists and has a local component. The batch mode run of the simulation

varied the number of neighbors included in the contagion effect. The results of these runs showed that the effect is not linear. Figure 7 reveals that density of neighbors changes the effects on property value but contrary to Immergluck and Smith's belief that their estimates were conservative due to areas outside of their study being denser [6], density is not a direct linear correlation to the contagion effect.

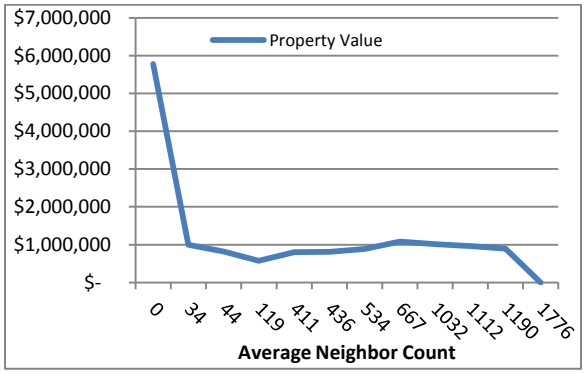


Figure 7: Average property values for selected neighbor counts.

Figure 7 is a sample of 10 random neighbor values plus the extremes, 0 and 1776. Though it is not scaled because of the random choices, it is apparent that the foreclosure contagion effect is affected by the number of neighbors included in the study confirming the fact that the effect is localized. The primary factor, however, is not the distance of a property but rather the number of properties in a given distance. This model will be repeated using different shape files with the same distances to show the effect of the neighborhood density on the foreclosure contagion.

VI. CONCLUSION

The property value effect of the spatial layout is one aspect to explore with ABMS. Additional analysis on areas such as the effects of natural and man-made boundaries and school and tax districts remains to be completed.

The idea that the number of neighbors available for producing an appraised value has implications for effects neighborhood design and planned community layouts could potential hold. There is the possibility that certain neighborhood designs can contain the foreclosure contagion effect.

REFERENCES

[1] Gangel, Marshall, Michael J. Seiler, and Andrew Collins, "Exploring the Foreclosure Contagion Effect Using Agent-Based Modeling", Journal of Real Estate Finance & Economics, Vol. 46, Issue 2, 2013.

[2] Genesove, D., and C. Mayer, "Loss Aversion and Seller Behavior: Evidence from the Housing Market", Quarterly Journal of Economics, 2001.

[3] Gilbert, Nigel. "Agent-Based Models", Series: Quantitative Applications in the Social Sciences, SAGE Publications, 2008.

[4] http://en.wikipedia.org/wiki/Hedonic_regression, accessed on 6 February 2013.

[5] Harding, John P. Eric Rosenblatt, Vincent W. Yao, "The Contagion Effect of Foreclosed Properties", JEL Classification, G12, G21, R31, 2009.

[6] Immergluck, Dan. Geoff Smith, "The External Costs of Foreclosures: The Impact of Single-Family Mortgage Foreclosures on Property Values", Housing Policy Debate, Vol. 17, Issue 1, 2006.

[7] Harding, J.P., E. Rosenblatt, and V. W. Yao, "The foreclosure discount: Myth or reality?," *Journal of Urban Economics*, vol. 71, no. 2, pp. 204–218, Mar. 2012.

[8] Kobie, Sugie, and Timothy F. Lee. "The Spatial-Temporal Impact of Residential Foreclosures on Single-Family Residential Property Values." Urban Affairs Review 47, No. 1, 2011.

[9] North, Michael J., Charles M. Macal, "Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation", Oxford University Press, 2007.

[10] Osland, Liv. "An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling." Journal of Real Estate Research 32, No. 3, 2010.

[11] Wilhelmsson, M. "Spatial Models in Real Estate Economics," Housing, Theory & Society, Vol. 19, No. 2, 2002.

Popularity or Proclivity? Understanding the Impacts of Instrumental and Intrinsic Preferential Attachment on Social Network Formation Employing Agent-based Modeling

Xiaotian Wang
Modeling, Simulation and Visualization Engineering Department
Old Dominion University
xwang009@odu.edu

Advised by Andrew Collins

Abstract

Barabasi-Albert model (BA model) is the classical algorithm used to describe the emergent mechanism of a category of networks, namely, scale-free network. This type of networks has been tested to approximately abstract the features of amounts of real world networks, such as, internet and World Wide Web. In this study, the author is trying to employ the simulation method—agent-based modeling—for learning the network behavior. It is argued that most of prior studies are rarely taking heterogeneity of agents into the analyzing the formation of networks. In fact, in the real world networks, people making choice to link with each other not only concern the degree of other people, but also care about intrinsic characters of each other. Hence, the author proposes an agent-based modeling method based on balancing weight between instrumental and intrinsic preferential attachment within the networks. The simulation result proves that power (define a person with more links having more power) is distributed evenly if people weight intrinsic preferential attachment more. There are hardly super hubs existing in such kind of networks. In contrast, power is owned by monopolies if people weight instrumental preferential attachment more.

Keywords:
Social networks, Scale free networks, Heterogeneity, Preferential attachment, Agent-based modeling.

1 Introduction

Social network analysis (SNA) has an especially long tradition in the social science. In recent years, a dramatically increased visibility of SNA, however, is owed to recent interests of statistical physicists, who start to employ statistical mechanics for the analysis of the formation of large-scale network. Barabasi-Albert model (BA model), addressed by Barabasi and Albert in 1999, is the first attempt to theorize this kind of phenomenon in the real world networks, that is, the degree distribution of networks can be described using power-law distribution [1]. Amount of large-scale real networks are studied and tested having the scale-free properties, include World Wide Web, internet, movie actor collaboration networks, etc. [2-4]. The main concern of this model is the impact of preferential attachment on network formation [5]. BA model assumes that nodes with more links (i.e., "popular nodes") are more likely to be connected when new nodes entered a system [1]. More interesting, BA model found that, as mentioned before, preferential attachment in a growing network leads to a power law degree distribution [1, 5]. The result is in sharp contradiction to the exponential distribution when nodes were randomly attached.

However, in many social networks, significant deviations from scale free behavior have been reported [6]. Accordingly, many variants of BA model are developed, and the main focus of these models is to reproduce the growth process of real networks by exploring the dynamic mechanisms at the node level (i.e., "popular" nodes are more likely to be connected). Despite of many new modifications, most of these models still share the key assumption that nodes with more links were more likely to be connected.

However, I found this line of research 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) [7]. 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 preferential attachment, I believe, is particularly consequential on such social networks as political communication.

In this study, I propose an integrative agent-based model of preferential attachment encompassing both instrumental calculation and intrinsic similarity. Particularly, it emphasizes the ways in which agent-heterogeneity affects social network formation. This integrative approach, I believe can strongly advance our understanding about the formation of various networks.

2 Integrative Models

There are two critical assumptions underlying BA model (i.e., the scale-free network) [5]. First, the network continuously expands by the addition of new vertices that are connected to the vertices already present in the system. Second, the probability that two vertices are connected is neither random nor uniform; instead, there is a higher probability that it will be linked to a vertex that already has a large number of connections.

At the agent-level, the scale-free network suggests that agents (i.e. vertexes) are instrumentally motivated to be connected to agents that are "popular." To incorporate this type of preferential attachment, we can assume that the probability Π that a new vertex $j + 1$ will be connected to vertex i depends on the connectivity k_i of that vertex [1], so that

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

BA model assumes that all the agents are homogenous, sharing the same instrumental preference of connecting with the popular. Although this homogenous instrumental preference attachment is a useful in describing micro dynamics of many social networks, it is often compromised by agents' inherent proclivity.

To capture the impact of agent heterogeneity on network formation, I propose another two assumptions in addition to those of the scale-free network. First, vertexes are intrinsically different from each other on certain aspects. A convenient example is party affiliation. In the United States, people can be labeled as Democrats and Republicans. Second, the probability that two vertices are connected is not solely determined by the connectivity of the existing vertexes. It is also determined by the instinct similarity between vertexes. Specifically, there is a higher probability that a new vertex will be link to a vertex that share similar characteristics. In other words, a new Democrat, when entering a social network, is prone to forming a connection with another Democrat but not a Republican. At the agent-level, my theorization suggests that agents are intrinsically rational, which implies that they are motivated to select parties that are characteristically proximate. This in term can be translated into Euclidean distance between vertexes.

$$\mathrm{P}(C_{N+1}, C_i) = \frac{1}{N-1}\left[1 - \frac{(C_i - C_{N+1})^2}{\sum_N (C_N - C_{N+1})^2}\right]^1$$

Where C_i represents the characteristic position of vertex i, C_{N+1} represents the characteristic

[1] $\sum_j \mathrm{P}(C_{j+1}, C_i) = 1$

position of a new vertex $N + 1$ that is about to join the existing network. A smaller Euclidean distance translates into more utility and hence contributes to the likelihood that the new vertex will be attached to vertex i. We can characterize this as intrinsic (or heterogeneous) preferential attachment in contrast to Barabasi and Albert's instrumental (or homogenous) preferential attachment.

Finally, we can update the probability U that a new vertex $N + 1$ will be connected to vertex i depends on the connectivity k_i of that vertex as follows,

$$U(\Pi, \mathrm{P}) = f(k_i, C_i, C_{N+1}) = \frac{\lambda k_i}{\sum_N k_N} + \frac{1-\lambda}{N-1} \cdot \left[1 - \frac{(C_i - C_{N+1})^2}{\sum_N (C_N - C_{N+1})^2}\right]$$

where λ is relative weight of instrumental preference, U is a product of Π, the instrumental preferential probability, and P, that intrinsic preferential probability. In light of this, BA model is only a special case of this integrative model (λ=1)

Therefore, each new node makes m new edges to remain in the network. Rather than solely attracted to the popular nodes, new nodes also weigh the extent to which other nodes are similar to themselves. The relative weights of this mixed preference are captured by λ. It is apparent that, $\lambda = 1$ will give rise to a purely rationality-driven social network (which is classical scale free networks), and $\lambda = 0$ will result in a value-driven social network. Above steps could be summarized as follows:

1. Start with two random value nodes link.
2. A random value new node is staring to consider about joining in the networks.
3. The original two nodes are being calculated by the probability defined above to be linked.
4. As we stressed in the section above, new node is considering about the weight between popularity and proclivity, calculate the probability for each node existing in the networks and link to a node with bigger probability value with random.
5. The fourth new node is going to join the networks, and three other existing nodes' probabilities are calculated to be decide if they are chose to be linked or not.
6. Loop from step 3.
7. Simulation is stopped when there are 502 nodes (in fact, the process repeat 500 times) in the networks.

The following schematic example is for translating the model into real world instance. A new person living in a community has to building up a network relationship with other ones in the same community. At the point of which person is who he wants to link with, he is concerning a balance between two features of other people: "Is he or she popular?" and "Is he or she a person more like me?" Relying on the result of his balancing, he makes a choice and links himself in

this community network. Meanwhile, he is going to be one alternative choice and to be evaluated by others if there are new persons join in this community. The model outcome proves that super hubs are generated if people only care about "I want to link myself to a more popular people, in other words, people with more resources". Another finding is that, as we can infer, scale free networks is a special case for $\lambda = 1$, when we add heterogeneity into the integrative networks.

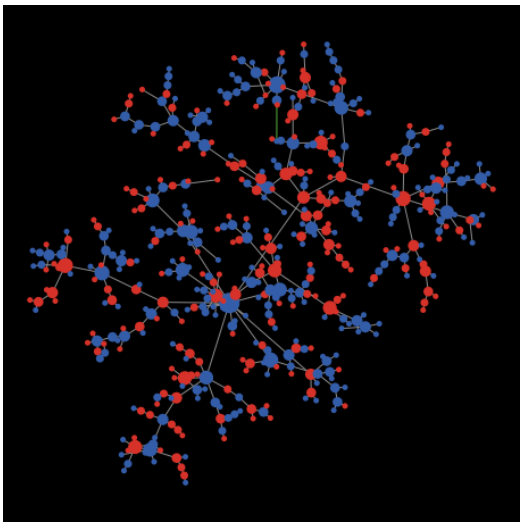
3 Simulation Result

The main purpose of this paper is to study how the intention from people affecting the structure of scale free networks. Based on the modified model introduced in the section 2, I simulate the networks model respective to different values of people's intention rate "λ", where "λ" indicate the objective choice from people between linking to the one with more links or the one "more like me". I set five different values to λ variable, and they are listed and explained in the table below:

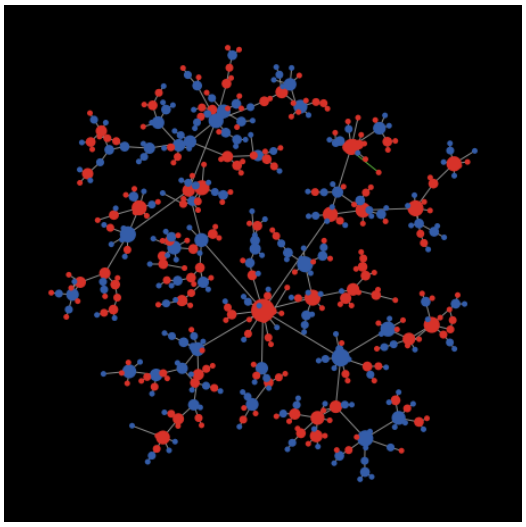
Table 1.

***λ* values**	**Categories**
0	People only concern about "Are you more like me?" or "Are we in the same party?"
0.25	People concern more about "Are you in my side", and also care about "Do you have more links" a little bit.
0.5	People concern these two parameters in the same level.
0.75	People concern more about "Do you have more links", and also care about "Are you in my side" a little bit.
1	People only care about "How many links do you have?"

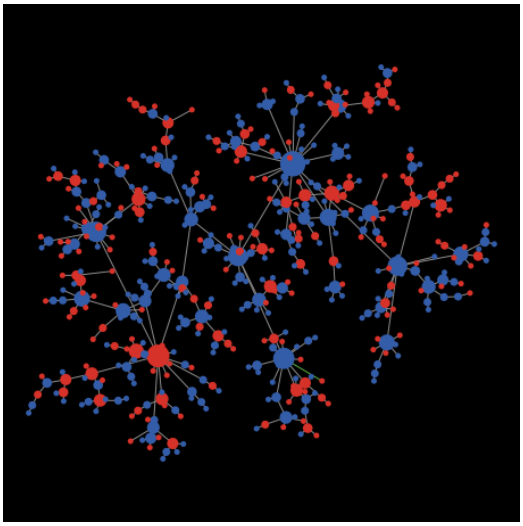
The network structures are generated regarding to different intention rate values based on the integrative model. Illustrated in the figure 1, as we can observe from the network structures of each graph, the network structures are demonstrated visually that they are getting more clustering as the value of λ increase, and the size of the hubs is getting bigger, explaining it in another way, the power (or resource) is distributed unevenly. There are 500 nodes in the networks in each graph.



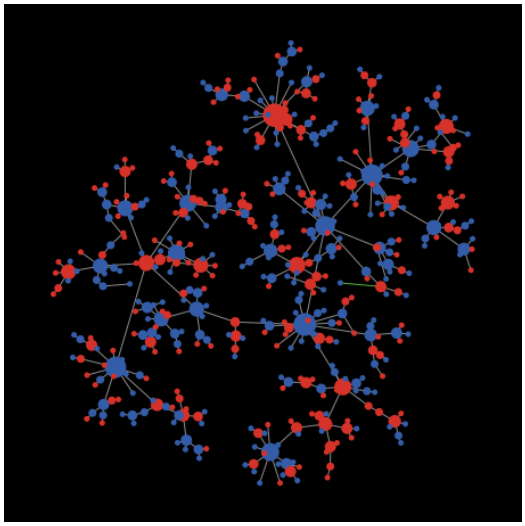
a) $\lambda = 0$



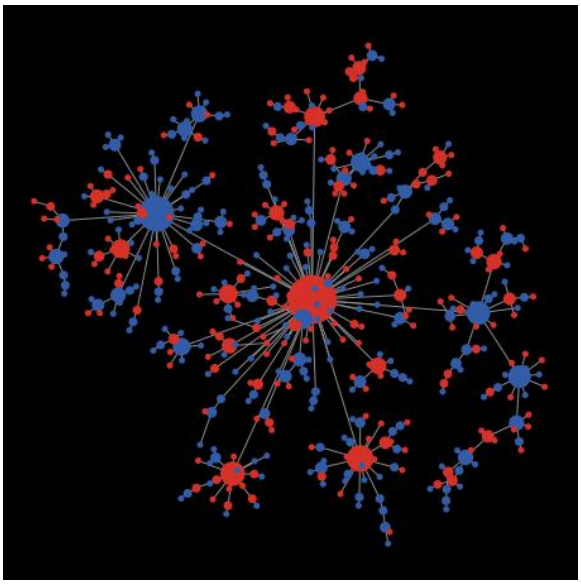
b) $\lambda = 0.25$



c) $\lambda = 0.5$



d) $\lambda = 0.75$



e) λ = 1

Figure 1. Networks topology respective to variant λ value

4 Conclusion and Discussion

In this paper, the author presents an integrative framework to understanding the impacts of preferential attachment in the emergence of social networks. In this integrative model, I try to stress that when joining a real social network, people are solely driven by instrumental calculation of connecting with the popular, as stated in the BA model. They are 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 author proposes an agent-based modeling method to simulate how agents balance between instrumental and intrinsic preferential attachment in forming their social connections. The simulation results suggest that power (define a person with more links having more power) is distributed evenly if people weight intrinsic preferential attachment more. There are hardly super hubs existing in such kind of networks. In contrast, power is owned by monopolies if people weight instrumental preferential attachment more.

References

[1] BARABASI A., Albert R, 1999, Emergence of scaling in random networks. *Science* 286: 509-512.

[2] Albert, R., H. Jeong, and A.-L. Baraba´ si, 1999, Nature (London), 401, 130.

[3] Faloutsos, M., P. Faloutsos, and C. Faloutsos, 1999, Comput. Commun. Rev. 29, 251.

[4] Watts, D. J., and S. H. Strogatz, 1998, Nature (London) 393, 440.

[5] ALBERT R., Barabasi A, 2002, Statistical mechanics of complex networks. *Rev. Mod. Phys*. 74: 47-97.

[6] Amirhossein Shirazi, Ali Namaki, Amir Ahmad roohi and Gholamreza Jafari, 2013, Transparency Effect in the Emergence of Monopolies in Social Networks, *Journal of Artificial Societies and Social Simulation* 16 (1) 1

[7] G. Caldarelli, A. Capocci , P. De Los Rios, and M. A. Mu˜noz, Scale-free Networks without Growth or Preferential Attachment: Good get Richer, http://arxiv.org/pdf/cond-mat/0207366v2.pdf. (Is this okay?)

C11EO Exploratory Model

Investigating Theoretical Frontiers in International Relations

Callie Davis
Old Dominion University
GPIS
Norfolk, VA
cdavi001@odu.edu

Abstract— C11EO is an agent-based model of a Collective Being. This feature drives C11EO's overall structure, performance and observation. C11EO explores possibilities for simulating flows of cognitive frameworks in a dynamic heterogeneous network. The simulation is stochastic: using random variables to achieve complexity for autonomous agents and the dynamic milieu of their interaction. In order to maintain complexity, a requisite for theoretical purposes, the system employs "diversity tactics" in multiple dimensions. C11EO may provide more interesting (pattern-rich) insight about globalization for the field of International Relations than similar models at present. Importantly, further development is required before implementation.

Keywords—agent-based; Collective Being; globalization; International Relations Theory; network

I. INTRODUCTION

Pervasive concepts of hierarchical control and assumptions of rational behavior dominate mainstream International Relations (IR) theory in the United States. This paper examines how an exploratory, agent-based model of social interaction may facilitate meaningful innovation for IR theorists and progress the state of the art in Modeling & Simulation. C11EO is a model of a Collective Being. This feature drives C11EO's overall structure, performance and observation. C11EO explores possibilities for simulating flows of cognitive frameworks in a dynamic heterogeneous network composed of autonomous agents. The simulation is stochastic: using random variables to achieve complexity for autonomous agents *and* the dynamic milieu of their interaction. In order to maintain complexity, a requisite for theoretical purposes, the system employs "diversity tactics" in multiple dimensions. C11EO may provide more interesting (pattern-rich) insight about globalization for the field of International Relations than similar models at present which rely with limited success upon rational choice theories to direct agent behavior within hierarchical (often networked) social environments. Importantly, further development is required before implementation of C11EO.

II. THEORETICAL GROUNDS FOR C11EO

A. Collective Being

C11EO is structured in the spirit of complexity theorist Yaneer Bar-Yam's description of human civilization as an emerging superorganism. The stochastic processes and parameters are guided by methods described by Gianfranco Minati and Eliano Pessa in an instructional text covering the development and management of a class of complex systems with special properties germane to our exploration.

The fidelity of the model depends upon the validity of C11EO's interpretation as Bar-Yam's description of civilization-as-organism, as well as the validity of Bar-Yam's claim that communications structures for social leverage are changing in correlation with advances in human civilization [2]. Minati and Pessa provide a method to analyze and authenticate the higher-order complex system described by Bar-Yam, and other sources [3], [4], [5] support many like concepts regarding the creation and specification of complex systems similar to C11EO, such as agent and environmental attributes, nature of inputs and outputs, and network characteristics.

B. Network Structure

C11EO's structure is a dynamic, heterogeneous network of autonomous agents [3]. Agents are created, linked, and set into motion at the beginning of each simulation. Because the nature of the arcs (depending upon relative values of incident vertices) can change over time, as well as the status of agents themselves, the network is considered dynamic. Heterogeneity of the network refers to the diverse ways agents are interlinked. Some agents will have more or less linkages, resulting in network structure that more closely resembles social reality than models which assume all agents influence all other agents; or models using grid or lattice structures which assume agents operate within confined neighbor groupings [3].

III. MODEL SPECIFICATIONS

C11EO simulates a Collective Being (CB). Methods for observing and analyzing this kind of system are extensively detailed in [1] and [2]. Common guidelines to achieve and sustain social complexity are found in [1], [2], and [3]. For the

most part, complexity is sustained by embedding a proclivity for diversity within different dimensions of C11EO's apparatus. Because methods and guidelines are available for the study of CBs, C11EO's performance is falsifiable for every run of the simulation. If the simulations fail to demonstrate requisite benchmarks for status as a Collective Being, then the results may be disregarded. The success rate of achieving Collective Being status, for what length of time, and from which initial conditions, if any, remain unknown until C11EO can be implemented. The exploratory value of the model is predicated upon successful simulation of the Collective Being environment.

A. *Vertices – Agents*

C11EO's vertices are autonomous agents. Their state is measured as part of a dynamic whole in discrete steps over time. Agents are capable of sending or receiving one "object" at a time. The object sent and received is always a representation of the sender's Cognitive Framework (CoF), which is also the effective identity of each agent. Received CoFs are processed internally at each vertex (by each agent). Agent processors take advantage of information sources both internal and external to the receiver. After processing a CoF, the receiving agent may or may not choose to change its own CoF. If the CoF changes, so then does the identification state of that vertex. After each whole-system decision cycle, the prevailing CoF, or identity message, is sent forth from each vertex.

Matrix-form profiles of characteristics that influence agent behavior as both target and sender of information over time, containing both discrete and continuous measures for variables, are employed to approximate the world view (CoF) of each agent [3]. These profiles are like baskets of ideas. The number of ideas per basket, the method of assortment, and the nature of observation implied (such as differences between continuous variables of attitude or dichotomous variables of behavior) is flexible.

C11EO's network structure will reveal some agents with greater prestige or influence than incident agents. These influential states can emerge naturally as a result of fertile clusters of receptive agents [4] —suggesting legitimacy of leaders—and can dissolve in much the same fashion as CoFs change over time. This organic emergence of leadership is a special feature of C11EO in contrast to controlled hierarchical models [5].

B. *Arcs – Interactions*

C11EO'S Arcs are interactions between agents. Forming the directional edges between vertices, arcs reveal patterns of connectivity within the Collective Being. The flow of information is affected by the nature of the relationship between vertices, which is dynamic as agents change over time. CoFs are objects of exchange and exploration: an adequate number of simulation runs should give experimenters a wealth of insight about the way people might exhibit and exchange various world views within the bounds of a Collective Being.

C. *C11EO's Baskets*

- Matrix formed of user-specified data.
- Serve several functions: sign of agent identity; message of sending agent; part of each agent's internal knowledge source for processing whether or not to adopt any or all of the contents of a received "basket".
- Always contains at least three variables, randomly assigned at the beginning of each simulation, pertaining to the agent's "sending" behaviors, "receiving" behaviors, and a threshold for susceptibility to environmental events which cue mass response for all agents in the same direction. These variables represent internal and external influences for each agent in a single-system, interdependent network.
- Users can manipulate the contents of baskets to suit various purposes of exploration, including adding variables or changing the nature of their measure.
- The contents of baskets may be tracked for comparison across different social contexts or network configurations, including various centrality characteristics, which could help us understand what kinds of interactions are most effective under specific circumstances.

D. *Diversity Drivers*

Sustainability of the CB system is linked to diversity of agents and interactions [1], [2], [3]. C11EO leverages critical diversity in multiple dimensions.

- Networks (vertices and arcs) and baskets are created with stochastic methods. Parameters for distribution of vertices and nature of arcs can be adjusted.
- Agents make decisions according to internal and external sources of information.
- Agents interact with other agents, and with an over-arching environment: different levels of analysis.
- The frequency and timing of interactions can be variegated.
- In an interdependent system, agents are both sources and receivers of influence… suggesting the possible formation of feedback loops or other interesting sub-network structures which can result in unpredictable system dynamics.

IV. Evaluation of C11EO in Context

C11EO's format as a model of alternative global reality makes it unique. The merits of modeling the globalizing world as a Collective Being, as compared to hierarchies of mainstream IR Theory, are perhaps best demonstrated by development and implementation of C11EO to the extent of creating observations which can be compared to those of mainstream models for accuracy in predictive or explanatory ability of real events. Short of implementation, C11EO is anticipated to compare favorably against other models in terms of both micro-level (individual) analyses and macro-level

(group) analyses. However, the exploratory nature of C11EO may not provide appropriate grounds for comparison in many cases.

While a literary review of IR theory and the state of the art for Modeling and Simulation are beyond the scope of this brief paper, the following distinctions can be made about C11EO:

- Some similar models manage the functionally inverse relationship between agent depth (individual detail) and the number of agents that can be supported within a simulation environment by imposing various hierarchies or using multiples of the same agent. C11EO is potentially capable of creating *millions* of agents with enough depth to capture individual complexity. This is in contrast to some of the best models currently in use, which may have fewer than 20 agents in a single simulation [4], [5].
- C11EO does not require leaps of faith regarding agent ability to process several hundred alternatives at any given moment, unlike models which rely heavily upon game-theoretical and individual psychological approaches to agent definition. C11EO agents are programmed with gentle overall preferences for behaviors, such as those which support homophily, that are primitive and ubiquitous in nature.
- C11EO is still under development and has not reached the implementation phase. Therefore it is difficult to do more than speculate as to the model's efficacy.
- C11EO is uniquely situated as an exploratory model because the CB performance restraints (non-CB sims are discarded) manifest as a form of validation: successful models are simulating levels of complexity seen only in reality; without the need for compensation in terms of reduced numbers of agents, use of cloned agents or unnaturally imposed hierarchies, or reduced complexity of individual agents.

ACKNOWLEDGMENT

C11EO has its seed in a past group project that constructed a discreet event simulation of evolving ideology. Ideas of team members Matthew Haase and Justin Pearl, both M&S graduate students at Old Dominion University, influenced the origins of C11EO.

REFERENCES

[1] G. Minati and E. Pessa. *Collective Beings*. New York, NY: Springer Science+Business Media, 2006.

[2] Y. Bar-Yam. *Dynamics of Complex Systems*. Reading, MA: Addison-Wesley, 1997. Ch. 9.

[3] W. A. Mason, F. R. Conrey, and E. R. Smith, "Situating social influence processes: Dynamic, multidirectional flows of influence within social networks," Pers. Soc. Psychol. Rev., vol. 11, 2007. pp 279-300.

[4] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. New York, NY: Cambridge University Press, 2011.

[5] B. G. Silverman, G. K. Bharathy, B. Nye, G. J. Kim, M. Roddy, and M. Poe. "M&S Methodologies: A Systems Approach to the Social Sciences." In *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*, J. A. Sokolowski and C. M. Banks, Eds. Hoboken, NJ: John Wiley & Sons, 2010, pp. 227-270.

The Induction of Community Dynamics in Online Social Networks

David Wright
ODU, MSVE Department
dwrig032@odu.edu

INTRODUCTION

The purpose of this research is to describe a new method for modeling communities, based on the words they have used in social networks, and for inducing community dynamics, such as merging, splitting, growing, and shrinking. The method constructs a multi-level hypergraph, in two levels: the first level represents the users and their social network; the second level represents the word network, the nodes of which are words and the edges are co-occurrences of words within a specific temporal or spatial range in a certain domain, e.g., Facebook. The users and words are gleaned from various sources, including Twitter, Wikipedia, and Stack Overflow, LiveJournal, Memetracker, DBLP, and Facebook. The applications of this method are manifold: link prediction, information cascades, community detection, item recommendation, user evaluation and trend prediction.

The method focuses on the way in which words are used. It does not purport to identify opinion, but rather attempts to predict based on the combined pattern of word use and historical data. For this, machine learning will be used to determine relevant coefficients, extracted from adjacency matrices, the central data structure for graphs. These in turn will be used to construct various conditional probability functions.

The method aims to answer any questions. For example, what is probability of re-tweeting conditioned on how many new edges will be formed in the user's word network, as well as the kind of new edges? What is the probability of a current admin A voting for an admin candidate B on Wikipedia based on the overlap of models of A and B? What is the probability of a new meme diffusing based on who sees it and the aggregate model of all users combined? Does user similarity fade with graph distance, or are there disjoint pockets of similar communities—doppelcommunities--and individuals? Which intermediate users must be friended by a group in order for that group to migrate across the network in order to merge with a target group?

BACKGROUND

User Similarity

Previous work in similarity is vast. [1] focus on user characteristics and the interplay between relative status and topic similarity as it applies to user evaluation. They evaluate user status based on a user's recent activity level: number of edits (Wikipedia), the number of questions asked or answered (Stack Overflow), and the number of ratings given (Epinions). They evaluate similarity as the overlap of user content and links. Their work shows that there are definite trends and models of user behavior within a social network based on similarity and status. The main conclusion is that there is a positive correlation between the user similarity and action (evaluation, rating, or voting).

Community Similarity

Community membership has been studied by [3], who describe the growth and evolution of groups in terms of the structural components of the network--the connectedness of friends—and in terms of the topics discussed among those friends. [4] discuss the topics of interest of communities, and conclude that networks emerge from the communication of users within a community, and that communities can be characterized or identified by the use of certain high-frequency words specific to the community.

Community Dynamics

There are six main types of community dynamics: growth and contraction, merging and splitting, birth and death. Research [3,4,8] has gone into analyzing the size and lifetimes of these communities, but none have researched how to induce those dynamics.

Active Friending

[10] gives an algorithm that optimizes an intermediate subset of nodes (and paths between them) between a source and a target, through which the source may become friends with the target--it is assumed that the distance between the source and the target is greater than 1. The algorithm works for one source and one target. Their future work will focus on one source trying to befriend multiple targets simultaneously with an algorithm that minimizes redundant work--probably dynamic programming.

CURRENT WORK

Network Model

This research aims to expand the latter notion [4] to include not just sets of frequently used words, but rather communities of words within the larger networks of words used by all users. This expansion results in a multi-level hypergraph.. One layer of the graph will represent the users, embedded in their social network; the other layer represents the word network formed by the aggregated words mined from user communication. Both graphs

Statistics

Various measurements must be gleaned from the data. This paper hypothesizes that there will be a certain threshold for the number of word communities that a user community actively uses, i.e., there will be a maximum level of entropy to the information passing between community members due to the limited time and interest of users. Data must be gathered on the correlation between highly mobile users—those who, more often than not, move in and out of communities—and the word communities they inhabit. There should be a parallel to the work on information pathways in communication networks [7].

Inducing Community Dynamics

This research hopes to expand the work of [5,6,10] to work for a group of source nodes and a group of target nodes, i.e., communities--groups with high cliquishness—who wish to migrate across the network will be given an optimal or near optimal set of intermediate. By expanding [10], algorithms can be derived to induce the dynamics above. For example, if two groups were to merge, and some distance on the graph separated them, there should be an optimal choice of intermediate nodes through which the group should navigate to bring about the merging. The counterpart to merging is splitting, and the algorithm should be essentially an inverse of the merging: which links to break to engender a split.

FUTURE WORK

The research consists of three main parts, of which each contains two types of mirror dynamics: growing and shrinking; merging and splitting; and birth and death. Data from online repositories will be used, as well as real-time data being collected by the author. Validation across domains (biological, social, information) will determine the fitness of the model and its predictive validity.

REFERENCES

[1] Anderson et al. Effects of User Similarity in Social Media... 2012

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg. Group Formation in Large Social Networks: Membership, Growth, and Evolution. *KDD '06*. August 20-23, 2006.

[4] 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.

[5] W. Chen, Y. Wang, S. Yang. Efficient Influence Maximization in Social Networks. *KDD '09*. June 28-July 1, 2009.

[6] W. Chen, Y. Wang, S. Yang. Scalable Influence Maximization for Prevalent Viral marketing in Large-Scale Social Networks. *KDD '10*. Juky 25-28, 2010.

[7] G. Kossinets, J. Kleinberg, D. Watts. The Structure of Information Pathways in a Social Communication Network. *KDD '08*. August 24-27, 2008.

[8] R. Kumar, J. Novak, A. Tomkins. Structure and Evolution of Online Social Networks. *KDD '06*. August 20-23, 2006.

[9] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

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

[11] D. Yang, H. Hung, W. Lee, W. Chen. Maximizing Acceptance Probability for Active Friending in On-Line Social Networks. arXiv: 1302.7025v1. 27 Feb 2013.

A Conceptual Model Utilizing Evolutionary Game Theory to Explore the Effects of Discount Factors on Interactions Between International Organizations

Rebecca Law, Old Dominion University

In my experience as a student and professional, I often am surprised (and excited) to see theory apply directly real world situation. So often there are striking degrees of variance in real world scenarios that lesson my belief in usefulness of theories. Never was this so apparent until taking an introductory international relations theory course and game theory course. As a sat through each of these courses, I recall trying to always apply one course's material to the other. At times this was difficult at best, but when the concept of shadow of the future was introduced to me in my international relations theory course and mentioned again in my game theory course along with discount factor I was elated to see this link.

The following paper presents a conceptual model utilizing evolutionary game theory to explore the effects of discount factor on interactions between international organizations.

Institutional Theory in International Relations

According to institutional theorists, international institutions matter because they perform four major functions. First, they reduce transaction costs of doing business because meeting through the apparatus of the institution reduces the individual cost for each state. Second, institutions allow for greater transparency and greater verification. For example, states are now privy to the capabilities of others. Finally, institutions increase cooperation by providing a standardization of communication and creating basic norms within the institution.

Shadow of the Future

The concept of the shadow of the future relates to how long the actors believe that they will be interacting. To expand on this premise, presume the actors are in a satiation where they will be interacting for a long time. In this case, the shadow of the future is characterized as "long." Conversely, if actors believe their interaction will be quick and brief, the shadow of the future is characterized as "short." The crux of the shadow of the future is about expectations.

International Relations Institutionalism and the Shadow of the Future

Do institutions lengthen the shadow of the future? Do institutions change how states view one another in their relationship? How?

Institutional theorists contend that institutions do in fact lengthen the shadow of the future. First, most institutions require states to sign publicly. This provides a shared audience of witnesses to the mutual agreement under which states enter institutions. Second, institutions provide an iterative success document, which in turn makes the actors believe

that there is architecture for their new relationship. From this, one can infer that the longer the shadow of the future, the increased likelihood of cooperation. But why? If the shadow of the future is long then states believe that the relationship will be a prolonged relationship. This belief has direct implications for a state's behavior now. A state must carefully consider their behavior and the consequences of cheating now or cooperating now. A further examination of this last statement is necessary.

Let us contemplate a situation within an institutional setting where The shadow of the future is long. States must consider that if they cheat now on an agreement they may be cheated on or punished later. Now, imagine a state in an institutional setting where the shadow of the future is short. States must consider that if they you cooperate now then they may be rewarded. Nevertheless, states can also get away cheating without the prospect of consequences in this scenario due to the brevity of the relationship.

Contrariwise, if a state chooses to cooperate in a short shadow of the future relationship the brevity of that relationship may not warrant time for reward either. The critical point here is that the longer the shadow of the future the more the cooperation now.

Game Theory and Shadow of the Future

Until now the concept of the shadow of the future has been discussed in the context of IR theory, specifically through the lens of an institutional theorist; however, the notion of the shadow of the future transitions seamlessly into the domain of game theory.

The concept of a discount factor- how much a decider prefers rewards now over rewards later- is the key to this transition. The discount factor is represented by the symbol δ, $with\ 0 < \delta < 1$.[1]

Hypothesis

From the knowledge gained on the concept of the shadow of the future, institutional theory and fundamental game theory applications such as δ, the author of this research paper proposes the following hypothesis for further examination:

If states enter an institution under the supposition that the shadow of future will be short and the shadow of the future in reality ends up being long, the effectiveness of that institution to mitigate a problem will be greatly diminished. Conversely, if states enter an institution under the supposition that the shadow of the future will be long and the shadow ends up being short in reality, the effectiveness of that institution to mitigate that problem will not be as diminished as severely as the first scenario.

Conceptual Model Development

The concept for model development is straightforward but requires some imagination on the reader's part. Imagine a problem exists that is global in nature. (Global warming or pollution is one such example.) The magnitude of the problem has become so great that it has now warranted multiple states and institutions enter an institution, or agreement, to address this growing global threat.

For the purpose of this game set up, however, the conceptual model focus is captured at the time when the

agreement or institution is formed. Furthermore, only two fictional institutions are considered.

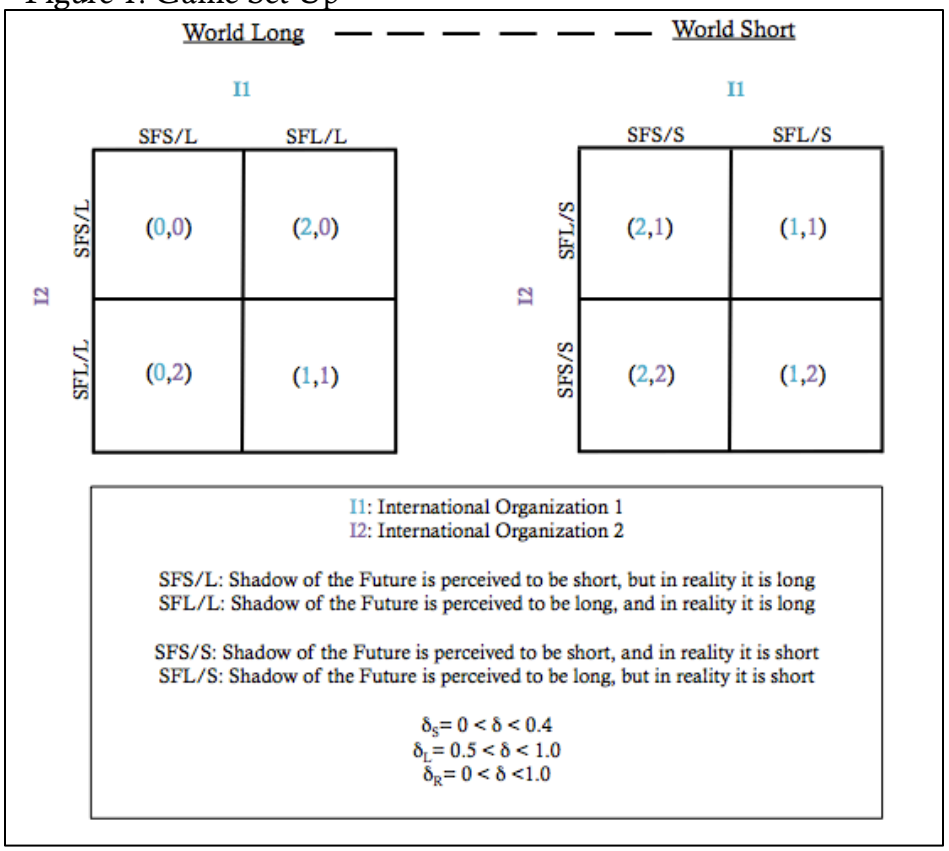
First, two fictional institutions represent the players of the game. The institutions shall be referred to as Institution 1 (I1) and Institution 2 (I2) throughout the rest of the paper.

Second, two distinct states of the world exist, initially determined by nature. World Long is a world in which the nature of the problem the institutions have joined warrants both I1 and I2 to anticipate a long shadow of the future. World Short is a word in which the nature of the problem the institutions have joined warrants both I1 and I2 to presume a short shadow of the future. Within these two states of the world, all possible combinations of I1 and I2's assumptions about the shadow of the future are represented. (see Figure 1.)

In addition, a indefinite Prisoner's Dilemma set up is chosen since interactions between I1 and I2 will be repeated although the state of the world will dictate how many times. A modification of the game to accommodate real world conditions is discussed later in the paper.

Next, various values for δ were assigned to I1 and I2 depending on what their assumption about the shadow of the future is. (see Figure 1.)

Figure 1. Game Set Up



Finally, the payoff structure for each player varies as it depends on which state of the world they exist.

How to Solve the Game

Although the game was not solved at the time this paper was written, further information regarding evolutionary game theory and Grimm Trigger strategy were studied to further refine this model. The following steps should be taken in order to solve this game:

- Calculate the expected payoff of each strategy in each state of the world.
- Determine the probability of each type of player.

- Assume a Grimm Trigger Strategy played by the single player who is incorrect of their knowledge of the state of the world. (i.e player playing SFS/L or SFL/S)
- Calculate the expected utility of each payoff.

Hypothesized Results

Preliminary results are abstract rather than concrete. The author postulates that the closer institutions' values for δ are to each other and δ_R, the greater the ability of these institutions to mitigate a problem. This is perhaps the reason why institutions are better able to increase cooperation. When expectations are similar, the likelihood of cooperation is increased as well as the ability to sustain that cooperation over a long period of time.

Conversely, when the institutions' values for δ are least compatible both with each other and δ_R, the ability of the institutions to mitigate a problem is impacted at its' greatest level.

Implications of Hypothesized Results

The reality of this situation requires an understanding of other factors that may influence the model. For example, a state or institution playing the strategy SFS/L may free ride (off of another state or institution playing SFL/L. In this case, the values associated within each payoff become questionable. For example, an institution that free rides would gain greater utility from free riding because they are not expending as much money and resources on the problem as the other institution; conversely, the institution that is the victim of free riding may in fact gain utility in knowing their intuition is mitigating the problem regardless of the fact that they are doing so at greater expense.

Case Study

The author of this study finds it useful to provide an example of where such research might benefit real world application. As such, a case study drawing upon current events involving maritime piracy off the coast of Somalia and the international community's attempts at mitigating this problem is discussed.

Background

Maritime piracy has existed since the seas were piled for trade. Piracy has traditionally been romanticized extensively by both writers and filmmakers, and has otherwise been consigned to the past. Recently, however, there has been an extraordinary increase in the number of attacks on commercial seafarers and their vessels. Modern-day piracy, specifically in the Horn of Africa (HOA) and Gulf of Aden, has transformed into a business-like industry incentivized by a high reward and low risk. Should this problem remain rampant, the impacts could have disastrous economic, political, and security consequences globally.

Current international counter-piracy measures represent an unprecedented level of international maritime collaboration that has emerged in a relatively short period of time, yet they have singularly failed to dent the incidence and scale of Somali piracy. Several reasons account for this.

First, existing strategies have been compartmentalized by the various agencies involved, which have largely not moved to develop, much less

implement, a uniformed set of goals and objectives.

Second, certain key stakeholders have been conspicuously absent in the formation of comprehensive approaches – notably ship owners who for cost-related and self-interested reasons have unwittingly made their assets highly vulnerable to attack.

Third, there are presently no agreed measures of effectiveness (MOEs) to determine the "success" and cost-utility (or otherwise) of policies designed to enhance maritime security off the HoA.

Fourth, and perhaps most fundamentally, the basic thrust of the response has been premised on a containment strategy that seeks to confront piracy at its end point – at sea – rather than at its root – on land.

It is the belief of this author that states and institutions seeking to eradicate maritime piracy in 2004 underestimated the will and desperation of the Somali pirates. As a result, the international community thought a show of naval force off the coast of Somalia would quickly address this problem. However, today the maritime piracy problem still remains rampant and is a booming economic industry not only off the coast of Somalia but in West Africa as well.

The author believes that states and institutions presumed the shadow of the future for anti-piracy coalitions would be short upon entering the initial agreements and coalitions. This is the reasoning for the overwhelming military response was initiated. However, over time the tactics and resolve of the Somali pirates changed becoming more innovative, sophisticated, and adaptive in their new environment.

Anti-piracy coalitions too adjusted their techniques due to learning. This can be seen by the advancement of attacks from sea to land where pirate camps conduct operations. There is also more consideration for land based approaches to piracy rather than a sea-based approach.

In conclusion, the nature of the Somali pirate problem required anti-piracy coalitions to enter agreements with the understanding that the shadow of the future would be long; however, the author believes this was not the case. As a result, piracy remains as vibrant as ever and anti-piracy coalitions are trying to keep pace.

Limitations

Recognition of limitations in a research project is just as critical as the results of the study. Limitations provide the author with an understanding of how the models can be improved for future research, as well as provide insights into gaps within the existing literature. Limitations to this study include, but are not limited to: time, beginner knowledge of game theoretic models and generalizations.

Time is always a critical dictator in conducting research. The nature and scope of this problem is vast. While the model is imperfect and incomplete, its usefulness is not diminished.

Second, the game theoretic models seem to be pliable. In other words, there was no definitive way to set up and execute the model. The author had a difficult time in choosing what she felt had been the "best" model.

Finally, a number of assumptions and generalizations were made in this study. For example, the game set up only took into account two players, when in reality there are a multitude of players with varying interests, expectations, and utilities that must be considered in order for the model to have greater fidelity.

Future Research and Applications

The author of this study presented the results in an abstract matter. Given a greater amount of time, it would be useful to perform analytical calculations.

However, the usefulness and repurposability of this model to other scenarios presents itself as a strength of this conceptual model and should not be weakened due to the lack of analytical work. This model could be applied to other problems of interest such as budgetary negotiations in the United States Congress for example.

Conclusion

The hypothesized results are preliminary at best, but hopefully have been thought provoking.

As with any research the author conducts, the author tends to walk away with more questions than answers. Questions for consideration include:

- What do the results of this study mean for institutions?
- Are institutions useful? To what extent? Under what conditions?
- Should it be required that intuitions be required to reveal their expectations about the shadow of the future in a signed document upon entering an institution?

It is the belief of the author that a generally accepted assumption about institutions exists. Because institutions increase cooperation, decrease transaction costs and provide some form of standardization and social norms, expectations of institution members are therefore uniform. This assumption is not only fundamentally wrong but also costly. A greater understanding of these concepts is necessary as the world's problems become broader, more complex, and critical.

References

1. Morrow, James D., "Game Theory for Political Scientists," Princeton University Press, 1994.

Reducing Stakeholder Fatigue: Integrating Governing Bodies, Committees and Working Groups Contribution to Risk Mitigation

David W. Flanagan, Andrew Collins, Ph. D., Barry Ezell, Ph. D.

***Abstract*— Hampton Roads Conducts about 286 emergency management meetings annually. The frequency of the meetings create economic and personnel strain on the region. Data was collected from 315 meetings minutes. After analyzing the data it was found that the region spends 2.1 million dollars and 34 thousand man hours annually. Also, more time is spent on travel to meetings then the actually meeting themselves. The conclusions of the project was that to reduce region fatigue utilizing more technology such as Video Teleconferencing (VTC) should be implemented at the various sites rather than having emergency managers physically attend meetings.**

INTRODUCTION

Meetings have away to become large and unbounded. Based on feedback from key personnel in Hampton Roads Planning District Commission (HRPDC), Hampton Roads (HR) Stakeholders are pulled in numerous directions that result in wasted time and unnecessary redundancies and duplicative processes. Over the past several years during the implementation of the Urban Area Working Group (UAWG), numerous capability assessments and follow-on related projects such as the development of regional strategy documents, sustainment plans and in general, overseeing the programmatic investments of the different grant funds, have significantly impacted the key stakeholder in the region and in many cases, participants outside the region. The impact of the problem is a perceived dilution of senior leader time [1]. Across the region, the number of man-hours lost and the associate cost can have an overall impact on mission. This project constructed a model to assess the meeting for all participants in the region as they relate to risk management and planning to understand the regional impact of man-hours and cost associated with meetings. The study began with the attempt to identify all meetings, associated tasks, purpose, attendees by name and affiliation, meeting cycle, inputs, outputs, products, and linkages to regional risk impact [4]. Not all the information was available but with the data collected a good baseline model of the regional burden was found.

I. METHODOLOGY

The methodology was designed to accomplish three tasks:

- Collection of information about current yearly meetings and exercise demand for HR risk managers and present that information in a summary format.
- Construct a cost model for those meetings and exercises. Cost is given in both time and financial requirement.
- Analyze the cost information and present summary, interesting observations and insights.

The meeting information provided here is given in time and monetary cost. The analysis method and data gaps found meant that several assumptions needed to be made within the analysis. Where possible, we aired on the side of caution, i.e., choose an assumption that would reduce overall cost (see table 3 in the appendix for details). As assumptions had to be made it is unlikely that all the data used in this analysis is correct and it is believed that monetary values given in this report are an underestimation of the true cost.

A. Data Collection

There were 25 meeting groups and eight exercises that were identified as appropriate for this analysis of those that meeting/minute data were provided. Of the 25 meetings, seven meetings contained major incompletions in the data sets, of which three were unusable for analysis purposes. The remaining data set was split into two groups:

- Data set "A" (18 groups): groups in this data set generally contained all meeting information required for analysis (Frequency, duration, and location) and enough information about the attendees to determine their required information (location and wage). Summary details are given in table 4 which is found in Appendix A.
- Data set "B" (4 groups and 8 exercises): these groups and exercises did not contain enough information to obtain a complete picture of their attendees.

The analysis was split over these two data sets. Where needed, dataset "A" information was used to make missing data judgments for dataset "B," i.e., number of attendees, duration etc. Even with the most complete meeting group datasets contained missing data. For dataset "A," this missing data was mainly information about the attendees. For the analysis that was conducted in this report, we need to know the attendee's annual wage and their home or normal business location. This information was used to determine the travel- times to-and-from meetings and monetary cost of the meetings; travel-time cost were given in terms of

individual attendees per hour wage rate and the standard Virginia Per Diem travel rate ($0.565). The required meeting group information was location, duration and frequency. Using this information, combined with the attendees per hour wage rate information, we were able construct a costing model and determine the expenses per meeting, in terms of attendees' wages. No overhead cost were included i.e., cost of meetings rooms, refreshments, etc.

1. Location

The location information was required to determine the travel costs of attendees to the meetings/exercises. The locations of all meetings and exercises were supplied; however, locations of attendees were not. Attendee's locations were found through internet searches, using sources like LinkedIn accounts, company websites, area codes from phone numbers, white pages and/or other appropriate websites. There was no guarantee that all the attendee information that we collected was up to date or correct. Where possible, the individual's profession was used to determine if they were the appropriate person for the meeting under consideration. When multiple people appeared in searched we chose the individual with the Hampton Roads location. To unify the analysis we amalgamated all the locations in one city to one central city location. There were a few exceptions, e.g., locations at military bases, HRPDC, Cox Communications and Virginia Natural Gas. Centralizing the locations does affect the accuracy of the result; this was considered acceptable given the other assumptions used. Once location data for meetings and attendees was complete we were able to determine travel time and distance for the attendee's traveling to and from a meeting location. This information was found using Google Maps (maps.google.com). It was assumed that there were no travel delays from incidents or traffic. The travel distance was used to determine the travel cost for the attendee and $0.565 per mile; it was assumed that all attendees travel by their own car. The travel time was also used to determine the wage cost of an attendee traveling to and from the meeting. When multiple locations were given for a meeting group, we assumed that the group's meetings where evenly split across those locations.

2. Wage

Two main sources for determining the wage information of the attendees: The Richmond Times-dispatch online for state employees, and The Virginia Pilot online for HR city employees. The Richmond times was from 2011 and only contained employees with salary over the state average of $52,559. The Virginia Pilot database was also from 2011 and categorized employees by job title, as opposed to name. From the two databases, only about 12% of the attendee's wages were discovered. The remaining attendees were assumed to have the state employee average. This figure was deemed a "low ball" estimate because it would be expected that highly qualified (and paid) professionals were those that attended the meetings/exercises. The wage that an individual gets is not the cost that company pays to keep them on staff; there is also fringe benefits (pension, medical insurance, etc.) and indirect cost (office heating and rental, administrative staff, managerial oversight cost, etc.). We assumed these values took the state average of 30% for fringe benefits and 53% for indirect cost. Given the lack of wage information for the attendees, this average salary formed the basis of the sensitivity analysis.

3. Missing Data

The meeting data provided contained lots of missing data needed for the intended analysis. Some data was universally absent, e.g., wage information, and some was only missing from certain meetings, e.g., meeting frequency. Assumptions had to be made to fill in the missing data gaps. When no information was given about a meetings location we assumed it was the same as the previous meetings in that group. When no information was given about the meetings duration, we assumed it was two hours long, which is the median of the durations of the other meetings we did not have information about. Dataset "B" contains meetings, with incomplete data, and exercises data. There was no information about who the attendees were for this data set so an average wage of $53k per year and the average travel cost from dataset "A," of $125 per person per meeting, were taken for the attendees of the meetings/exercises of dataset "B." For the four meeting groups with incomplete attendee data, it was assumed that the average of 25 attendees attended those meetings and the meetings occurred monthly.

B. Results

The results from the model are in four parts: summary statistics, group information, travel information, and some sensitivity analysis. The region holds about 286 meetings per year on HR emergency management matters; this figure includes about 30 meetings associated with four exercises per year. We had two years' worth of exercise data thus 315 meetings were considered in total. There were 22 meeting groups that regularly met and made up these meetings. The HR region spends about $2.1 M annually on emergency management and about 34,000 man-hours are annually spent at meetings including transportation of personal to-and-from them. One of the most interesting results found were that there seems to be very little overlap in attendees, which was unexpected: there were 515 attendees from a list of 402 people considered and only 77 people were involved in overlap. Of those that overlapped, three people attended five different groups each. There are various general statistics that can be drawn from the two datasets, shown in Table 1. Dataset "A" contained the bulk of the overall cost as it contains the most groups (18). It is noticeable that, per year, the amount of time and money spent of exercise (and preparation meetings) is substantially less than the amount spent on general meetings.

			Annual Totals			
	Attendees	Number	Man-hours	Travel Cost	Meeting Cost	Total Cost
Groups - Dataset A	515	18	24600	$796,000	$782,000	$1,578,000
Groups - Dataset B	100	4	4200	$150,000	$110,000	$260,000
Exercise 2011	699	4	3300	$87,000	$103,000	$190,000
Exercise 2012	1112	4	6600	$139,000	$225,000	$364,000
Total 2011	1314	26	32100	$1,033,000	$995,000	$2,028,000
Total 2012	1727	26	35400	$1,085,000	$1,117,000	$2,202,000

Table 1: Descriptive statistics from datasets

The split between travel expenses and actual meeting time expenses are about the same. For the general meeting groups, more is spent on travel, and for exercises more is spent on the actually meeting. This difference is because the exercise meetings tend to be substantially longer than the normal meeting groups. The meetings have variation in cost. This is due to the different travel times required for the different attendees at each meeting and from different durations of the meetings. Thus it is possible to have meeting with the same number of attendees with vastly different cost. About half of the man-hours allocated to meetings are spent on travel.

A regional map was used to show the travel requires for the regions. Figure 1 shows the amount of travel from outside the Hampton Roads region. The red dots indicate meeting locations and green are starting points. The thickness of the connecting line is indicative of the amount of people traveling from that location. From Figure 1, it is showed that a large number of people travel from Richmond to attend regional meetings. This has a large impact of meeting cost for the region. Figure 2 represents the Hampton Roads area. The amount of meetings held at a location is indicated by size of dot. The green are the starting locations, and the lines indicated the amount of people travelling. Chesapeake holds the most meetings per year, with Norfolk and Virginia Beach doing the most travel.

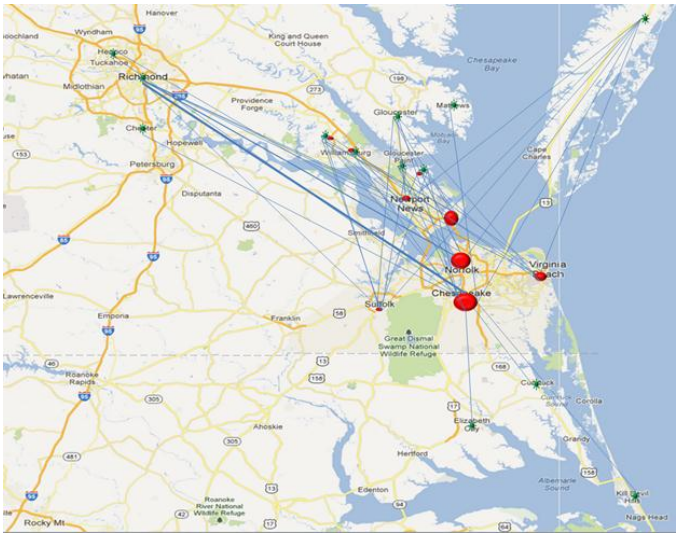


Figure 1: Regional View

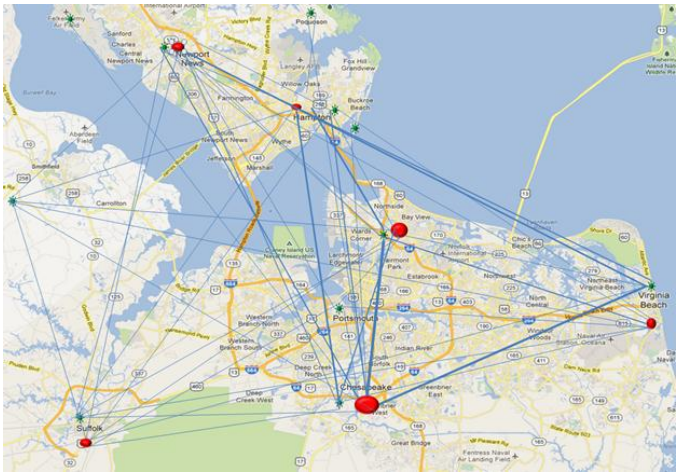


Figure 2: Large out of region view

Most (88%) of wages used in the analysis came from a single estimate; this estimate was of the state average of $53k. Given the large impact of this value on the analysis displayed, it seemed appropriate to conduct some sensitivity analysis on this value. The wage value was used for the attendees was varied in two ways:

- Low salary: every salary was assumed to have the state average of $53k per year
- High salary: every attendee was assumed to have the average salary of the attendee's salaries that we were able to find, i.e., 12% of all the attendees, $112k per year.

Salary	Travel cost	Meeting Cost	Total
Dataset A - meetings			
Original salary values	$796,000	$782,000	$1,578,000
High salary ($112K)	$1,234,000	$1,479,000	$2,713,000
Low salary ($53K)	$737,000	$700,000	$1,437,000
Dataset B - meetings			
Original salary values	$150,000	$110,000	$260,000
High salary ($112K)	$240,000	$236,000	$476,000
Low salary ($53K)	$143,000	$112,000	$255,000
Exercises 2012			
Original salary values	$139,000	$225,000	$364,000
High salary ($112K)	$222,400	$482,600	$705,000
Low salary ($53K)	$132,000	$229,000	$361,000

Table 2: Sensitivity analysis

The high salary nearly doubled the overall yearly cost; $3.9M.

II. Recommendations and Conclusion

The model and results provide a fresh perspective on the emergency management meeting load within Hampton Roads. The results indicate a lot of meeting expenditure and time goes on transportation between Richmond and Chesapeake. The region would benefit from utilizing more technology to hold meetings such as Video Teleconferencing (VTC). This would cut man hours nearly in half by eliminated all the hours spend on travel. The majority of meeting costs come from the regular meetings as opposed to exercise meetings. These results might seem counterintuitive because exercise cost will appear as a budget line item whereas regular meeting costs are hidden.

Thus regular meetings should be targeted for cost saving measure.

The Hampton Roads regions spends about $2.1M annually on emergency management planning and 34k man-hours are annually spent at meetings included exercises and transportation of personnel to them. These results are based of the 22 groups meeting frequently during a year and four exercises per year. The purpose of representing the meetings in monetary cost terms is that it is easier to sell a monetary saving than a time saving. That is easier to report that you wish to save "X" dollars in meetings and exercises requirements than "Y" hours of emergency manages time. The monetary value is not explicit saved because wages still need to be paid but by not attending excessive meetings, the emergency managers are able to spend time on their "day job" which might have otherwise been passed on other team member resulting in more staff at the emergency managers office location. We made no attempt to recommend meeting groups or exercises to be eliminated from the calendar; even if evaluating the purpose of each meeting was possible, it is unlikely that we would have the insight or expertise to see all purpose and functions of the groups. The unstated purposes of a meeting might be to disseminate information around different agencies, the open evaluation of ideas by attendees, etc. The unstated purpose of the meetings might not all be positive and could have negative effects, which has been coined "organizational misbehavior" [2] Thus this research cannot make recommendations based on which meeting group is deemed unnecessary. This model and its results were hoped to be used as a catalyst for discussion with the Hampton Roads emergency management community.

APPENDIX

Assumption	Effect on Estimate
Non-inclusion of MMRS Strike, REMTAC Debris and HRICAC	⇩
Unknown salaries were assumed to be the average state salary of $53K	⇩
No fixed overhead costs were used for determining meeting costs	⇩
Attendees travelled by own car to and from meetings	⇩
When selecting attendee's location from multiple possibilities, we chose locations in Hampton Roads over any others.	⇩
There were no travel delays from incidents or rush-hour traffic.	⇩
Used 2011 wage information	⇩
All attendees came to every meeting	⇧
Assume each meeting in a group has the same duration and same location	⇧
Amalgamated many meeting locations within a city to single location.	-
Locations for Cox Communications and Virginia Natural Gas were assumed to be in Norfolk.	-
Meeting duration was the average duration of the times that were presented in data.	-
People locations were assumed from online sources such as linkedin accounts, white pages, area codes in phone numbers, or their employer's webpage.	-
Starting locations and distance to meeting were taken from central points in cities. Except for Cox Com, VA Nat. Gas and HRPDC.	-

Table 3: Assumptions

Meeting Name	Attendees	Cost per meeting	Total cost per year
Area Maritime Security	20	$3,600	$43,000
Chesapeake Local Emergency Planning Committee (LEPC)	35	$5,700	$68,000
City Readiness Initiative Health Planner (CRI)	18	$3,400	$40,000
Hampton Road Planning District Commission (HRPDC)	45	$13,800	$165,000
Hampton Roads Regional Catastrophic Planning Team (HRRCPT)	92	$29,300	$351,000
Hampton Roads Highway Incident Management (HRHIM)	13	$4,600	$18,000
Hampton Roads Hurricane Evacuation Workgroup Meeting (HRHEW)	33	$8,600	$207,000
Hampton Roads Inmate Evacuation Planning (HRIEP)	41	$8,600	$103,000
Hampton Roads Emergency Management Committee (HREMC)	23	$5,200	$125,000
Metropolitan Medical Response Service (MMRS) Healthcare Committee	26	$6,700	$79,000
Metropolitan Medical Response Service (MMRS) Oversight Committee	25	$5,200	$21,000
Overlay Regional Operability Network (ORION) Steering Committee	23	$10,300	$123,000
Peninsula Local Emergency Planning Committee (LEPC)	17	$3,000	$12,000
Regional Emergency Management Technical Advisory Committee (REMTAC)	17	$4,100	$49,000
Regional Emergency Management Technical Advisory Committee (REMTAC) Special Needs	15	$2,200	$27,000
Regional Emergency Management Technical Advisory Committee (REMTAC) WebEOC	6	$1,000	$4,000
Tidewater Emergency Medical Services (TEMS) Council	23	$4,100	$49,000
Urban Area Working Group (UAWG)	43	$7,600	$91,000
		Grand Total	**$1,578,000**

Table 4: Cost per meeting and yearly cost of the different meeting groups from Dataset "A."

ACKNOWLEDGMENT

The project team would like to thank Robert Lawrence, HRPDC (the project sponsoring organization) for his tireless meeting data collections and information clarification, and Donna Brehm, CRA inc., for her collection of exercise data. Without either of them, the data analysis presented in this report would not have been possible.

REFERENCES

[1] Redick, James A. (2010) "Regional Disaster Planning: Observations and Recommendations, A Case Study Of Hampton Roads." City of Virginia Beach, Office of Emergency Management. W.K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Collins, Andrew J.; Joshua Behr; Lisa Blair; Saikou Diallo; Mikel Petty; Solomon Sherfey; John Sokolowski; Andreas Tolk; Charles Turnitsa; and Eric Weisel ; Modeling and Simulation Standards Study: Objects & Tools; Workshop Report; Suffolk, VA: VMASC, May 11- 13, 2011.

[3] Collins, Andrew Ph.D. (2013) "Reducing stakeholder fatigue: integrating Governing Bodies, Committees and Working Groups Contributing to Risk Mitigation," V13-001; Suffolk, VA: VMASC; 2013.

Determinants of seafarer fatal and non-fatal injuries in container vessel accidents

Yishu Zheng
Department of Modeling, Simulation and Visualization Engineering
Old Dominion University
yzhen003@odu.edu

Abstract: This paper investigates the determinants of the probability of fatal and non-fatal crew injuries in container vessels based upon individual accident report data collected by the US Coast Guard for the 2001- 2008 period. A Probit regression statistical model is used to uncover these determinants. Results suggest that: (1) a crew member is less likely to be injured in a containership and ro-ro container vessel accident if the vessel has steel hull and the accident occurred at night time; (2) a crew member is more likely to suffer from a fatal injury if fire is involved in the accident; (3) if both containership and ro-ro container are registered under US flag regulation, powered by diesel and built with steel hull, it is less probable that the crew member will suffer a fatal injury during an accident.

Keywords: Vessel injury, probability, Probit, containership[1], ro-ro container[2], container barge[3].

I. INTRODUCTION

Vessel accidents are incidents caused unintentionally that may result in ship and cargo damages, and worker accidents and deaths [1]. Although shipping safety regulations have greatly improved, key challenges including increased ship sizes and human factors are still important aspects in regulating shipping safety. During the last twenty years, ship size has increased significantly as it directly increases the volume of trade and reduces the transportation unit cost; this is especially true for long-haul sea trips. Although ship size has increased significantly compared, the number of crew member at a vessel did not increase at the same rate. Traditionally, a minimum of 13 seafarers were needed to operate a container ship. Now, the larger container vessels have between 19 and 34 crew members. From the statistics, between 1980 and 2000, accidents caused 0.15% to 0.55% ship losses each year, with an average of 0.3% a year. Human loss incidents also resulted, on average, in 700 crew members lost per year, which is an individual risk of 1.6% per ship per year [2].

Vessel safety regulations and enforcements have the objective to prevent and minimize the severity of vessel accidents. Maritime safety is governed by a combination of national and international rules. The International Maritime Organization (IMO) is presented as the main United Nations agency in charge of watching for the safety and security of shipping and the prevention of marine pollution. In 1998, the International Safety Management Code (ISM) was adopted internationally by the IMO country members, providing the international standards for the safe management and operation of all types of vessel. The ISM Code requires the shipping lines to document the management procedures in conformity with the provisions of the regulations in order to prevent, detect and minimize unsafe human behaviors. The motivation of ISM Code is that human factors are considered one of the main causes of vessel accidents [3]. For further discussion of human error in shipping, see [4-6]. Recently, the IMO has adopted a new Code of International Standards and Recommended Practices for a Safety Investigation into a Marine Casualty or Marine Incident (Casualty Investigation Code), which was introduced in January 2010 [7]. This is the first time that the IMO produced a standardized definition of a serious injury[4].

On the other hand, ship classification societies are also responsible for maintaining technical standards and enforcing vessel safety rules. Lloyd's Register, Bureau Veritas, and Det Norske Veritas are some of the main classification societies that inspect vessels to ensure their seaworthiness, verify the national-flag requirements, and regulate their adherence to the international safety rules. For more discussion of classification societies, see [8-9]. Some countries have established Port State Control (PSC) which helps to inspect foreign vessels of all flags that enter their ports. The Port State Control Officer will be in charge of the inspection with the purpose of verifying the conditions of the ship and equipment, the operating manners, and the competency of the crew members on board. For further discussion of Port State Control, see [10-12].

[1] Containerships are designed to carry cargoes loaded in truck-size containers, generally TEUs (Twenty-foot Equivalent Units) and FEUs (Forth-foot Equivalent Units).
[2] Roll-on/roll-off container vessel is designed to carry containers, trainers, general cargo, and wheeled cargo (trucks, trailers, cars, etc.)
[3] Container barge is a flat bottom boat designed to transport freight at river and canal waterways.

[4] "An injury which us sustained by a person, resulting in incapacitation where the person is unable to function normally for more than 72 hours, commencing within seven days from the date when the injury was suffered".

The purpose of this paper is to study the severity of container vessel accidents from the standpoint of the crew, i.e. to investigate determinants of the probability of fatal and non-fatal crew injuries and missing crew in container vessel accidents. The results can be used by flag and flag-of-convenience state and classification societies to evaluate the vessel safety programs and regulations for reducing the number of crew injuries in container vessel accidents, and to develop related cost-benefit analysis of the different safety-enhancing measures.

A recent research study analyzing fatality and injury incidents (with data obtained by the Lloyd's Register Educational Trust Research Unit from 2000-2005) was done by Seafarers International Research Center (SIRC), at Cardiff University [13]. The study revealed that cultural differences in risk and behavioral differences between national groups may cause a difference in nationalities in terms of crew injury rates.

Determinants of the fatal and non-fatal crew injuries of individual commercial US and foreign flag bulk, container and tanker vessel accidents (investigated by the US Guard for the time period 1981-1991) have been analyzed by Talley [14]. The results showed that the number of fatal crew injuries are greater: (1) for tankers than for container or bulk ships; (2) if the accident cause is human rather than environmental or ship related; (3) for fire/explosion than for collision, material/equipment failure or grounding accidents; and (4) for multiple ships than for single-ship collisions. The number of non-fatal injuries is also greater for human related accidents, fire/explosion, and multi-vessel accidents. In another study of the determinants of towboat vessel accidents (based upon US Guard investigations for the time period 1981-1991), McCarthy and Talley stated that the number of fatal and non-fatal crew injuries are greater for: (1) docked or moored vessels; and (2) fire/explosion accidents.

Casualty rates of flag states (with data of 121 flag states who suffered casualties to merchant vessels from 1997-1999 recorded in the Lloyd's casualty database) was investigated by Alderton and Winchester [15]. Results showed that: (1) Flag of Convenience (FOC) have a worse record than both second/international registers and national flags; (2) significant variables exits within the same FOC vessel group, (3) the newer and faster growing FOCs are most likely to have poorer safety records than the established ones due to profit regimes. In a more recent study, Talley et al [16] investigated the determinants of fatal and non-fatal injuries and missing crew members in freight ship, tanker and tugboat vessel accidents (with data obtained from the US Coast Guard for the time period of 1991-2001). Results indicates that: (1) freight ship and tanker non-fatal crew injuries are greater when the vessel is moored or docked, and when high winds and cold temperature exist; (2) tugboat non-fatal injuries are higher with poor visibility environments; (2) freight ship accident crew deaths are expected to increase with vessel age, tanker fatal injuries are higher for fire accidents, and tugboat fatal injuries are greater for capsize and lake accidents; (4) freight ship missing crew increases with vessel age, and tugboat missing crew increases with fire and lake accidents.

This study complements the literature on determinants of container vessel fatal and non-fatal crew injuries by estimating separate vessel-accident equations for containership, ro-ro container and container barge accidents. For this investigation, detailed data of individual container vessel accidents for the time period 2001-2008 is obtained by the US Coast Guard to perform the analysis. The study is organized as follows: a model of container vessel accident injured and deceased crew is presented in Section 2, followed by a discussion of the data in Section 3. Results from the estimations are presented in Section 4 and estimated marginal effects are discussed in Section 5. Conclusions are detailed in Section 6.

II. THE MODEL

The probability of crew injuries in a container vessel accident (PINJ) is expressed as a function of the number of crew on board (CREW), and the severity of the vessel accident (VSLDAM), i.e.

$$PINJ = f(CREW, VSLDAM) \quad (1)$$

A positive a priori relationship exists between CREW and PINJ, i.e. as the number of crew on board increases, the more likely it will become that a member of the crew will suffer a fatal or non-fatal injury in a vessel accident, ceteris paribus. The vessel damage severity (VSLDAM) is expected to have a non-negative effect on the total number of crew injuries (PINJ), given that a damaged vessel does not necessarily result in fatal or non-fatal injuries. Both variables, CREW and VSLDAM, are modeled as function of other variables. The number of crew on board (CREW) varies with vessel age (VSLAGE) and vessel size, which is measured in terms of gross tonnage (GROSSTON), i.e.

$$CREW = g_1(VSLAGE, GROSSTON) \quad (2)$$

The relationship between VSLAGE and CREW is expected to be positive, since older ships tend to be more labor intensive than newer ones. GROSSTON should have a non-negative effect on CREW, because a larger sized vessel does not necessarily require a larger crew size. The vessel damage severity (VSLDAM) is expected to vary depending on the type of container vessel (VTYPE), type of accidents (ATYPE), vessel characteristics (VCHAR), type of vessel propulsion (PTYPE), type of vessel hull construction and design (HTYPE), weather/visibility at time of accident (VISIBILITY), and vessel operation phase (OPHASE), i.e.

$$VSLDAM = g_2(VTYPE, ATYPE, VCHAR, PTYPE, HTYPE, VISIBILITY, OPHASE) \quad (3)$$

Accident reports of three types of container vessels are used in this study: container ship (CONTAINER_S), ro-ro container (RORO_CONTAINER), and container barge (CONTAINER_B). It is unclear which type of container vessel will incur greater accident damage. The type of vessel accidents may be an abandon (ABANDON), allusion

(ALLISION), capsize (CAPSIZE), collision (COLLSN), emergency response (EMRESP), explosion (EXPLODE), fire (FIRE), flooding (FLOODING), grounding (GROUNDING), loss of power (LSPOWER), loss of stability (LSSTAB), material or equipment failure (MATFAIL), sinking (SINKING), and loss of maneuverability (MANEUVER)[5]. The a priori relationship of each accident type with VSLDAM is indeterminate.

Vessel characteristics include vessel age (VSLAGE), ship size in terms of the gross tonnage (GROSSTON), and whether the vessel is a US flag vessel (USFLAG). A positive relationship is expected between VSLAGE and VSLDAM, since the design, material and construction technology of older vessels tend to be more susceptible to accidents. A negative relationship is expected between USFLAG and VSLDAM, since U.S. flag registry has some of the stricter ship safety regulations and standards among ship registries.

The type of vessel propulsion may be diesel propulsion (DIESENG), gasoline propulsion (GASENG), or turbine propulsion (TURBINE). The relationship of vessel propulsion with VSLDAM is unclear. The type of the vessel hull construction and design is distinguished by aluminum (ALUMHULL), fiberglass (GLASSHULL), steel (STEELHULL), wood (WOODHULL), double hull (DOUBLEHULL), double side (DOUBLESIDE), and double bottom (DOUBLEBOTM). The a priori relationship of vessel hull construction and VSLDAM is undetermined, since WOODHULL might suffer more damage in fire/explosion accidents, but less damage in grounding accidents. On the other hand, vessels with DOUBLEHULL, DOUBLESIDE and/or DOUBLEBOTM may have less damage than single hulls in collision, allision and grounding accidents.

Visibility at the moment of the accident is depends on whether the accident occurred during nighttime (NIGHT) or daytime. A positive relationship is expected between NIGHT and VSLDAM. The ship operation mode can be either adrift (ADRIFT) or not at the moment of the vessel accident, the relationship of ADRIFT and VSLDAM is unclear.

Replacing the variables in equations (2) and (3) by the variables used to classify or measure them and then substituting these equations into equation (1), the reduced-form equation is obtained as follow:

$$PINJ = F\,(CONTAINER\_S, RORO\_CONTAINER, CONTAINER\_B, ABANDON, ALLISION, CAPSIZE, COLLSN, EMRESP, EXPLODE, FIRE, FLOODING, GROUNDING, LSPOWER, LSSTAB, MATFAIL, SINKING, MANEUVER, GROSSTON, VSLAGE, USFLAG, DIESENG, GASENG, TURBINE, ALUMHULL, GLASHULL, STEELHULL, WOODHULL, DOUBLEHULL, DOUBLESIDE, DOUBLEBOTM, NIGHT, ADRIFT) \quad (4)$$

Equation (4) is estimated separately for the probability of non-fatal injuries (INJURY) and the probability of fatal injuries (DEATH) in a container vessel accident.

III. DATA[6]

Estimates of equation (4) are obtained utilizing detailed data of individual container vessel accidents of container ships, ro-ro containers and container barges that occurred during the 8-year time period 2001-2008. These data were obtained from the US Coast Guard Marine Safety Management System (MSMS) database. Table 1 (Appendix) presents the variables used in the estimation of equation (4), their specifications and descriptive statistics (mean and standard deviation) for all three types of container vessels. The mean statistics reveal that 34.5% of the incidents were caused by material/equipment failure, 22.8% by loss of maneuverability, 7.9% by allision, 5.4% by loss of power, 4.3% by grounding, 2.9% by collision, 2.8% by fire, 1.2% by flooding, and 0.2% by explosion and sinking. A large amount of accidents occurred at containership, followed by ro-ro container and container barge; therefore, descriptive statistics by each type of container vessel is presented in the next few columns of Table 1.

From the results obtained on Table 1, the mean statistics for the explanatory variables reveal that for container ship accidents, 34.4% was material failure, 24.4% was loss of maneuverability, 6.1% was allision, 6.1% was loss of power, 4.8% was grounding, 3.4% was collision and 2.4% was fire. For ro-ro containers accidents, 37.9% was material failure, 16% was loss of maneuverability, 12.1% was allision, 4.8% was fire and 2.4% loss of power. For container barge accidents, 55.5% was allision, 11.1% was grounding, 11.1% was loss of maneuverability, 5.55% was flooding and 5.55% was loss of stability.

Furthermore, the average age for container ships, ro-ro containers and container barges are 14.5, 20.3 and 24.7 years respectively. The average size of container ship, ro-ro container and container barge accidents are 31,903, 30,304 and 4,646 gross tons. In terms of hull construction, over 90% of container ships and ro-ro containers, and 70% of container barges were built with steel hull. The majority of the container ship and ro-ro container accidents, and all container barge accidents occurred while the vessel was underway. Finally, the mean statistics reveal that 41.3%,

[5] An abandoned ship accident occurs if the ship is abandoned by its crew members. If the ship strikes a stationary object, the accident is referred to as an allision. A collision occurs when a ship strikes or was struck by another ship when navigating on the water. Emergency response occurs if the U.S. Coast Guard provides assistance to the ship in the accident. For an explosion accident, an explosion is the cause of the ship accident. A fire accident occurs when the fire causes the vessel accident. For a flooding accident, flooding is the cause of the accident. In a grounding accident, the ship is in contact with the sea bottom or a bottom obstacle. A loss of power occurs when the cause of the vessel accident is a reduction in power supply. A material or equipment failure typically involves equipment failure on board the ship. A maneuverability accident occurs when a reduction in a ship's maneuvering capability is the cause of the vessel accident.

6 All the tables mentioned from this section are displayed in Appendix.

34.9% and 33.3% of container ship, ro-ro container and container barge accidents occurred at night, respectively.

IV. ESTIMATION RESULTS

The variables in Table 1 are utilized in the estimation of equation (4), i.e., in the estimation of the parameters of equation (4). Specifically, estimates of equation (4) can be found in Table 2 (Appendix) for which the dependent variables are the binary variables non-fatal injuries (INJURY), and fatal injuries (DEATH). The Probit regression statistical model is used to obtain the estimates, since Probit regression estimation restricts the prediction of the dependent binary variables to lie in the interval between zero and one. For further discussion of the Probit regression statistical model, see [17-18].

In the second columns of Table 2, the estimated parameters of equation 2 for the heretofore discussed explanatory variables for non-fatal crew injury can be found. In the third columns of Table 2, the estimated parameters of equation 2 for the heretofore discussed explanatory variables for fatal crew injury can be found. Container barge accidents are not discussed as the limited data was not able to produce statistically significant results.

From the non-fatal injury Coefficient results obtained from Table 2, for containership accidents, Chi-square statistic estimate is 68.42, and it exceeds the 13.28 critical value necessary for significance at the 0.01 level for 4 degrees of freedom. Results reveal that three of the hypothesized explanatory variables – DIESENG, STEELHULL and NIGHT – are statistically significant at the one (DIESENG and STEELHULL) and five (NIGHT) percent level. The negative signs for the three coefficients suggest that: an individual is less probable to be injured in a containership accident if the ship uses diesel propulsion and has a steel hull. Also, the probability of a crew member injured in a containership accident decreases if the accident occurred at night. For ro-ro container vessel accidents, Chi-square statistic estimate is 44.01, and it is above the 15.09 critical value necessary for significance at the 0.01 level for 5 degrees of freedom. Results reveal that four of the hypothesized explanatory variables are statistically significant at the five (MATFAIL) and one (GROSSTON, DIESENG and NIGHT) percent level. The negative signs for the coefficients suggest that: the probability that a crew member gets injured in a ro-ro container accident decreases if the accident involves material/equipment failure, the vessel has a large size, the vessel has diesel propulsion, or the accident occurred at night.

In Table 2, the container vessel accident fatal injury regression estimates are presented third column. For containership accidents, Chi-square statistic estimate is 29.98, and it exceeds the 11.34 critical value for significance at the 0.01 level for 3 degrees of freedom. Results reveal that four of the hypothesized explanatory variables, - FIRE, USFLAG, DIESENG and STEELHULL – are statistically significant at the one percent level. The negative signs for the three coefficients indicate that an individual is less probable to suffer from a fatal injury in a containership accident if: the vessel is under US flag registration, has diesel propulsion and a steel hull construction. For ro-ro container vessel accidents, the model is again a good fit since the chi-square statistic is large and statistically significant at the 0.01 level, well above the 15.09 critical value for 5 degrees of freedom. Results reveal that six of the hypothesized explanatory variables, – FIRE, GROSSTON, USFLAG, DIESENG, STEELHULL and NIGHT – are statistically significant at the one, five, ten, one, one and five percent level, respectively. The positive coefficient for FIRE suggests that an individual is more likely to suffer from a fatal injury on both containership and ro-ro containers if fire is involved in the accident.

V. MARGINAL PROBABILITIES

Although the signs of the estimated Probit coefficients suggest either an increase or decrease in the probability of a fatal and non-fatal injury in a container vessel accident, the coefficients themselves do not measure the correct marginal probability effects for nonzero observations of the dependent variable. However, estimates of the correct marginal probability effects can be derived using the estimated coefficients. For the detailed procedure, we refer the reader to Greene [17]. In Table 3 (Appendix), we have shown the resulting marginal probabilities.

The second column of Table 3 presents the marginal effects of non-fatal injuries. Steel hull has the largest marginal probability effect on an injury in a containership accident, followed by diesel engine and night time accidents. That is, when a container vessel is made of steel hull, the probability of non-fatal injuries in an accident is decreased by a factor of 0.10. For non-fatal injuries occurred at a ro-ro container, diesel engine has the largest marginal effect, followed by night time accidents, material failure and vessel size. E.g. when a ro-ro vessel is propelled by diesel, the probability of non-fatal injuries in an accident is decreased by a factor of 0.11.

The third column of Table 3 presents the marginal effects of fatal injuries. Fire accidents have the largest marginal probability effect on fatal injuries in a containership accident. That is, when the accident involves fire, the probability that an onboard crew will incur a fatal injury increases by a factor of 0.106. For fatal injuries occurred at a ro-ro container, fire accident has the largest marginal effect, followed by diesel propulsion, steel hull, night time accidents, US flag and gross tonnage.

VI. CONCLUSION

The goal of this study was to reveal the determinants of the probability that a crew member onboard a ship experiences a fatal or non-fatal injury in container vessel accidents. To this end, we have used accident data in the period 2001 to

2008 collected by the US Coast Guard. The data include information on injuries and deaths of crew members, type of container vessels, type of accident, vessel characteristics, type of vessel propulsion and vessel hull construction, visibility and the vessel operation phase at the time of accident. Equations for crew injuries and deaths were estimated using Probit regression technique. The determinants may be used by states and classification societies to evaluate the performance of container vessel safety programs. For example, to determine whether safety programs have an impact on the determinants of container vessel accident crew injuries and deaths, as found in this study.

The estimation results revealed that a crew member is less likely to be injured in a containership and ro-ro container accident if the vessel has a steel hull and the accident occurred at night time. The probability of crew injury in a ro-ro container accident decreases if the accident is caused by material and equipment failure. The estimation results also suggest that a crew member is more likely to suffer from a fatal injury if fire is involved in the accident. If both containership and ro-ro container are registered under US flag regulation, powered by diesel and built with a steel hull, it is less probable that the crew member will suffer from a fatal injury during an accident. For ro-ro container vessels, the larger the ship size, the less likely the crew member will suffer from a fatal or non-fatal injury.

The author expects to extend this current research in order to explore and discuss about the strengths and weaknesses of Probit methodology in investigating the determinants of crew fatal and non-fatal injuries. For future work, more studies and different approaches will be considered to address the limitations of the model.

REFERENCES

[1] Talley, W. K., 2008, Maritime Safety, Security and Piracy. The Grammenos Library Informa. London.

[2] FAULKNER, D., 2003, Time for action on shipping safety. *Lloyd's List*, 28 April, 6.

[3] MACRAE, C., 2009. Human factors at sea: common patterns of error in grounding and collisions. *Maritime Policy & Management*. **36**, 21-38

[4] GROSS, R., 1994, Safety in sea transport. Journal of Transport Economics and Policy, **28**, 99-110.

[5] MILLAR, I. C., 1980, The need for a structure policy towards reducing human-factor errors in marine accidents. *Maritime Policy & Management*, **6**, 9-15.

[6] ABRAMS, A., 1996, New rules put focus on human factors. *Journal of Commerce*, 2 May, 8B.

[7] INTERNATIONAL MARITIME ORGANIZATION, 2008, *Casualty-Related Matters: Code of the International Standards and Recommended Practices for a Safety Investigation into a Marine Casualty or Marine Incident*, MSC-MEPC.3/Circ.2 ref T1/12.01. Retrieved 15 March 2013, from http://imo.org

[8] BOISSON, P., 1994, Classification societies and safety at sea: Back to basics to prepare for the future. *Marine Policy*, **18**, 363-377.

[9] MILLER, M., 1998, Classification societies from the perspective of United States Law. Retrieved 12 March 2013, from http://k.b5z.net/i/u/2023483/i/Classification.pdf

[10] PAYOYO, P.B., 1993, Implementation of international conventions through port state control: an assessment. *Marine Policy*, **18**, 379-292.

[11] Talley, W. K., 2002, Maritime safety and accident analysis. In*: The Handbook of Maritime Economics and Business*, edited by C. Grammenos (London: Lloyds of London Press), pp. 426 – 442

[12] TALLEY, W. K., Forthcoming, Regulatory issues: the role of international maritime institutions. In: *Handbook of Transport Strategy, Policy and Institutions*, edited by D. A. Hensher and K. J. Button (Oxford: Pergamon).

[13] ELIS, N., BLOOR, M., and SAMPSON, H., 2010. Patterns of seafarer injuries. *Maritime Policy & Management*, **37**, 121-128

[14] TALLEY, W.K., 1999, The safety of sea transport: determinants of crew injuries. Applied Economics, **31**, 1365-1372.

[15] ALDERTON, T., and WINCHESTER, N., 2002, Flag states and safety: 1997-1999. *Maritime Policy and Management*. **29**, 151-162

[16] TALLEY, W.K., JIN, D., and JITE-POWELL, H., 2005, Determinants of crew injuries in vessel accidents. Maritime Policy & Management, **32**, 263-278.

[17] GREENE, W., 2012, Econometric Analysis, 7th edition, Prentice Hall, Upper Saddle River NJ.

[18] STOCK, J., WATSON, M., Econometrics, 2nd edition, Pearson Addison Wesley, Boston MA.

APPENDIX

Table 1. Variable definition and descriptive statistics

Variables	Description and measurement	Mean (Standard Deviation) TOTAL		CONTAINER_S		RORO_CONTAINER		CONTAINER_B	
Dependent variable									
INJURY	1 if non-fatal injury, 0 otherwise	0.172	(0.130)	0.014	(0.118)	0.034	(0.182)	0.000	(0.000)
DEATH	1 if fatal injury, 0 otherwise	0.004	(0.064)	0.005	(0.071)	0.000	(0.000)	0.000	(0.000)
Explanatory variables									
Time of vessel accident									
YEAR	Year 2001 -2008	2004	(1.732)	2004.051	(1.731)	2003.723	(1.654)	2004.389	(2.304)
Type of container vessel									
CONTAINER_S	1 if containership, 0 otherwise	0.816	(0.388)						
RORO_CONTAINER	1 if roll on/roll off container, 0 otherwise	0.169	(0.375)						
CONTAINER_B	1 if container barge, 0 otherwise	0.015	(0.121)						
Type of vessel accident									
ABANDON	1 if an abandonment, 0 otherwise	0.001	(0.029)	0.001	(0.032)	0.000	(0.000)	0.000	(0.000)
ALLISION	1 if an allision, 0 otherwise	0.079	(0.270)	0.061	(0.240)	0.121	(0.327)	0.555	(0.511)
COLLSN	1 if an collision, 0 otherwise	0.029	(0.170)	0.034	(0.182)	0.010	(0.098)	0.000	(0.000)
EMRESP	1 if an emergency response, 0 otherwise	0.011	(0.107)	0.011	(0.105)	0.014	(0.120)	0.000	(0.000)
EXPLODE	1 if an explosion, 0 otherwise	0.002	(0.040)	0.001	(0.032)	0.005	(0.070)	0.000	(0.000)
FIRE	1 if a fire, 0 otherwise	0.028	(0.165)	0.024	(0.154)	0.048	(0.215)	0.000	(0.000)
FLOODING	1 if a flooding, 0 otherwise	0.012	(0.110)	0.010	(0.010)	0.019	(0.138)	0.055	(0.236)
GROUNDING	1 if a grounding, 0 otherwise	0.043	(0.204)	0.048	(0.214)	0.014	(0.120)	0.111	(0.323)
LSPOWER	1 if a loss of power, 0 otherwise	0.054	(0.226)	0.061	(0.240)	0.024	(0.154)	0.000	(0.000)
LSSTAB	1 if a loss of stability, 0 otherwise	0.003	(0.057)	0.003	(0.055)	0.000	(0.000)	0.055	(0.236)
MATFAIL	1 if a material failure, 0 otherwise	0.345	(0.476)	0.344	(0.475)	0.379	(0.486)	0.000	(0.000)
SINKING	1 if a sinking, 0 otherwise	0.002	(0.040)	0.000	(0.000)	0.005	(0.070)	0.000	(0.000)
MANEUVER	1 if a loss of maneuverability, 0 otherwise	0.228	(0.419)	0.244	(0.429)	0.160	(0.368)	0.111	(0.323)
Vessel characteristics									
GROSSTON	Vessel size in gross tonnage (gross tons)	31,358	(19,569.53)	31,903.71	(19059.88)	30304.64	(21382.06)	4646.083	(2,140.41)
VSLAGE	Vessel age (years)	15.615	(9.038)	14.476	(8.561)	20.302	(9.605)	24.667	(2.708)
USGLAG	1 if a US flag vessel, 0 otherwise	0.411	(0.492)	0.371	(0.483)	0.553	(0.498)	0.944	(0.236)
Type of vessel propulsion									
DIESENG	1 if the vessel is under diesel propulsion, 0 otherwise	0.811	(0.392)	0.856	(0.351)	0.660	(0.475)	0.055	(0.236)
TURBINE	1 if the vessel is under turbine propulsion, 0 otherwise	0.113	(0.317)	0.078	(0.269)	0.291	(0.455)	0.000	(0.000)
Type of vessel hull construction									
ALUMHULL	1 if aluminum hull, 0 otherwise	0.002	(0.040)	0.000	(0.000)	0.010	(0.098)	0.000	(0.000)
STELHULL	1 if steel hull, 0 otherwise	0.895	(0.307)	0.893	(0.309)	0.917	(0.276)	0.7222	(0.461)
DOUBLEHULL	1 if double hull, 0 otherwise	0.005	(0.070)	0.006	(0.077)	0.000	(0.000)	0.000	(0.000)
DOUBLEBOTM	1 if double bottom, 0 otherwise	0.002	(0.050)	0.003	(0.055)	0.000	(0.000)	0.000	(0.000)
Visibility at time of accident									
NIGHT	1 if night time, 0 otherwise	0.401	(0.490)	0.413	(0.493)	0.349	(0.478)	0.333	(0.485)
Vessel operation phase									
ADRIFT	1 if the vessel sets adrift, 0 otherwise	0.004	(0.064)	0.004	(0.063)	0.005	(0.070)	0.000	(0.000)
Observations	Total number of observations	1217		993		206		18	

Table 2. Container vessel accident fatal and non-fatal injuries: Probit regression estimates

Variables	1. Non-fatal injury Coefficient (t-statistics)		2. Fatal injury Coefficient (t-statistics)	
	CONTAINER_S	RORO_CONTAINER	CONTAINER_S	RORO_CONTAINER
Explanatory variables				
Type of container vessel				
CONTAINER_S	-0.784 [a] (-4.87)		-1.144 [a] (-5.46)	
RORO_CONTAINER		-0.495 [b] (-2.27)		-13.873
Type of vessel accident				
MATFAIL		-0.35 [b] (-2.04)		
FIRE			1.513 [a] (3.08)	1.362 [a] (3.35)
Vessel characteristics				
GROSSTON		-0.0015 [a] (-3.57)		-0.001 [b] (-2.26)
USFLAG	-1.225 [a] (-3.02)	-0.63 [c] (-1.67)		
DIESENG	-0.566 [a] (-3.31)	-1.212 [a] (-8.18)	-1.046 [a] (-3.94)	-1.212 [a] (-4.81)
STELHULL	-1.09 [a] (-7.52)		-0.882 [a] (-2.97)	-1.014 [a] (-3.72)
Visibility at time of accident				
NIGHT	-0.529 [b] (-2.52)	-0.817 [a] (-3.65)		-0.715 [b] (-2.43)
Chi-square test	68.42	44.01	29.98	151.54
Observations	1199	1199	1199	1199

a) significant with p-value < 0.01
b) significant with p-value < 0.05
c) significant with p-value < 0.10

Table 3. Container vessel accident fatal and non-fatal injuries marginal effect

Variables	1. Non-fatal injuries marginal effect		2. Fatal injuries marginal effect	
	CONTAINER_S	RORO_CONTAINER	CONTAINER_S	RORO_CONTAINER
Explanatory variables				
Type of vessel accident				
MATFAIL		-0.02		
FIRE			0.106	0.084
Vessel characteristics				
GROSSTON		-0.009		-0.004
USFLAG			-0.026	-0.014
DIESENG	-0.035	-0.11	-0.04	-0.0525
STELHULL	-0.1		-0.032	-0.041
Visibility at time of accident				
NIGHT	-0.023	-0.039		-0.016