Testing the Waters with Undergraduates (If you lead students to HPC, they will drink)

Angela B. Shiflet
Department of Computer Science
Wofford College
Spartanburg, S. C. 29303 USA
001-864-597-4528
shifletab@wofford.edu

George W. Shiflet
Department of Biology
Wofford College
Spartanburg, S. C. 29303 USA
001-864-597-4625
shifletgw@wofford.edu

ABSTRACT
For the Blue Waters Undergraduate Petascale Education Program (NSF), we developed two computational science modules, "Biofilms: United They Stand, Divided They Colonize" and "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better." This paper describes the modules and details our experiences using them in three courses during the 2009-2010 academic year at Wofford College. These courses, from three programs, included students from several majors: biology, chemistry, computer science, mathematics, physics, and undecided. Each course was evaluated by the students and instructors, and many of their suggestions have already been incorporated into the modules.

Categories and Subject Descriptors
K.3.2 [**Computers and Education**]: Computer and Information Science Education - Computer Science Education, Curriculum

General Terms
Design, Experimentation, Measurement.

Keywords
Computational Science, High-Performance Computing, Educational Modules, Biofilms, Social Networks, Blue Waters, Undergraduate, Petascale.

1. INTRODUCTION
With NSF funding, the Blue Waters Undergraduate Petascale Education Program [1] is helping to prepare students and teachers to utilize high performance computing (HPC), particularly petascale computing, in computational science and engineering (CSE). UPEP supports three initiatives:

- *Professional Development Workshops* for undergraduate faculty
- *Research Experiences* for undergraduates
- *Materials Development* by undergraduate faculty for undergraduates

The goal of the Materials Development initiative is "to support undergraduate faculty in preparing a diverse community of students for petascale computing."

For this program, the authors developed and class tested two computational science modules, "Biofilms: United They Stand, Divided They Colonize" and "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better," which are available on the UPEP Curriculum Modules site [2]. This paper describes and discusses the modules and our experiences using them in the courses Modeling and Simulation, High Performance Computing, and Mathematical Modeling at Wofford College [3] during the 2009-2010 academic year.

Several of the students in these classes are obtaining Wofford's Emphasis in Computational Science (ECS). Bachelor of Science students may obtain an ECS by taking Calculus I, Introduction to Programming and Problem Solving (in Python), Data Structures (in Python and C++), Modeling and Simulation, and Data and Visualization and doing a summer internship involving computation in the sciences [4]. Meaningful applications that illustrate fundamental concepts and techniques, such as those in the above modules, are crucial in their computational science education.

2. MODULES
2.1 Pedagogy
Prerequisites are for the modules are minimal, requiring no programming or calculus background but the maturity to read the material. Students who used the modules ranged from first- to fourth-year with majors from biology, chemistry, physics, mathematics, computer science, and undecided. Both modules provide the biological background necessary to understand the applications, the mathematical background needed to develop

models, and references for further study. "Biofilms: United They Stand, Divided They Colonize" has ten (10) multi-part quick review questions with answers to provide immediate feedback, while "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better" has six (6) such questions. The former module provides twenty-three (23) project assignments, and the latter has nineteen (19) projects for further exploration.

To help with implementation, example solutions in various systems are available for download from the UPEP Curriculum Modules site [1]. Accompanying the biofilms module are a *Mathematica* implementation and high performance computing simulations for two and multiple processors in C/MPI along with an accompanying walkthrough of the code by student intern Shay M. Ellison. The epidemic module has simulations in *Mathematica* for a small dataset of 18 people and for a large dataset of 1000 people randomly selected from Network Dynamics and Science Simulation Laboratory (NDSL) synthetic data for the population of Portland, Oregon [5].

2.2 High Performance Computing in Modules

In line with the aims of UPEP, both modules have a section on "Computing Power" that discusses the need for high performance computing (HPC) within the contexts of the particular applications. The biofilms section on the topic points out that biofilms are highly complex with numerous features and the version in the module considers form, not function, in 2D as opposed to 3D. Thus, HPC is usually necessary for more involved and realistic biofilms models. The other module discusses how petascale computing is needed for processing large datasets involving millions of people and their activities and for more sophisticated computations, such as studying the nature of epidemics and the impacts of policy decisions on controlling epidemics in urban environments.

With each module, the "Computing Power" section can be covered for information only, as a starting point for class discussion, or as motivation for the students' own HPC project development. Moreover, students can develop sequential or high performance computing versions of many of the projects with some assignments requiring HPC and others asking for timing comparisons of parallel codes and their sequential counterparts.

2.3 Class Testing

At the end of the semester, students in Mathematical Modeling, which used the epidemics module, and High Performance Computing, which used both modules, were asked to complete a questionnaire about the modules, while students in Modeling and Simulation completed a general questionnaire about the course. The first author taught the latter two courses, while another professor taught Mathematical Modeling.

The questionnaires for the HPC class had the students rate the following statements from 1 (strongly disagree) to 5 (strongly agree):

- I understood the science applications in the module.
- I understood the mathematics in the module.
- I understood the algorithms in the module.
- The module was readable.
- The program helped me understand parallel processing with MPI.
- My team/I could complete the C/MPI program.

They were also asked to elaborate about the above scores, particularly those below 4, to indicate what they like best and what they found most difficult about the module, to give corrections and suggestions for improvement, and to list under what circumstances would they anticipate that high performance computing would be useful in modeling that application.

The questionnaire for the Mathematical Modeling class, which did not use HPC, included the first four questions above along with the following 1-5 rated questions:

- The modeling assignment(s) helped me understand the material.
- My team/I could complete the modeling assignment(s).

They were asked similar discussion questions to those for the HPC class as well as the modeling project(s) they did related to this module and under what circumstances would they anticipate that high performance computing would be useful in modeling social networks. The professor completed a similar questionnaire and responded to a follow-up email about the module. Unfortunately, because the questionnaire was distributed at the end of the semester, only one Mathematical Modeling student and the professor completed the questionnaire.

Results of the questionnaires are described below.

3. MODULE 1: BIOFILMS

3.1 Scientific Question

A biofilm is a community of very small organisms that adhere to a surface (substratum) in an aqueous environment [6]. Examples of this ubiquitous phenomenon are dental plaque, the sticky substance covering the breathing passages of cystic fibrosis patients, antibiotic resistant bacterial colonies, and the microbial film in wastewater treatment. Because these communities have such important impacts, scientists are seeking to understand better the structure and function of biofilms [7], and the application provides good motivation for computational science students in courses such as High Performance Computing and Modeling and Simulation.

3.2 Computational Models and Algorithms

The module developed a cellular automaton simulation in two dimensions (2D) of the formation of the structure of a biofilm without regard to its function. Each discrete time step of the simulation contained the following phases:

- Diffusion of nutrients
- Growth and death of microbes
- Consumption of nutrients by microbes

The module covers the basics of cellular automaton simulations including boundary conditions and models for diffusion, biofilm growth model, and nutrient consumption. Afterwards, the material develops generic algorithms for the models, a simulation program, and a visualization.

3.3 Assessment of Simulation's Results

A series of module figures shows several steps of a visualization of bacteria grids and associated nutrient grids, such as in Figure 1 of this article. Using the figures, a section on "Rubric for Assessment" examines empirically successes and shortcomings of this simulation of a structural formation of a biofilm.

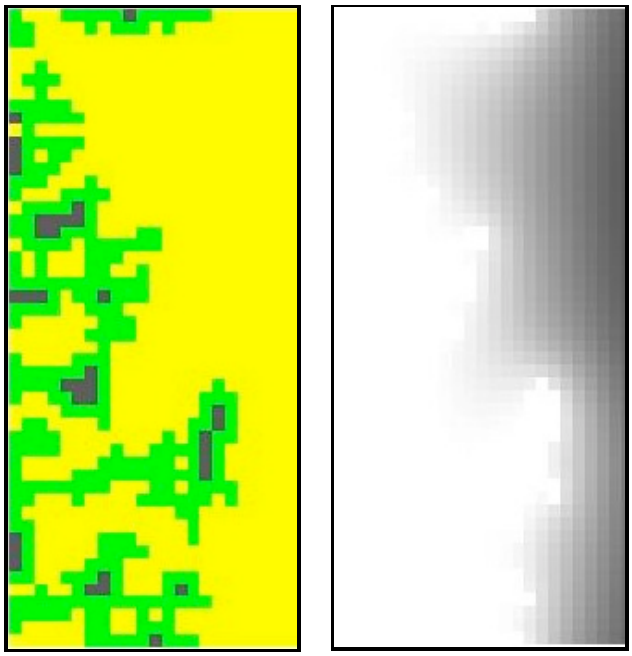

Figure 1. Simulated biofilm with corresponding nutrient grid

3.4 Class Testing in Modeling and Simulation

With *Introduction to Computational Science: Modeling and Simulation for the Sciences* [8] as a text, half of Wofford's Modeling and Simulation course (COSC/MATH 201) covers system dynamics modeling using a tool, such as *STELLA*®, *Vensim*®, or *Berkeley Madonna*®. The second half of the semester employs a computational tool, such as *Mathematica*®, *Maple*®, or *MATLAB*®, for developing cellular automata (CA) simulations, where the world under consideration consists of a rectangular grid of cells and each cell has a state that can change with time according to rules. We start the second part of the semester by considering random walk simulations and having projects on the formation of polymers and crystals. Additional important CA algorithms involve diffusion and spreading, and simulation of the formation of biofilm structures provides a significant application of these techniques.

With the background provided by this application and these concepts, over a two-and-a-half week period, the four students in the class (three biology majors and a mathematics major, all ECS students; a freshman, a junior, and two seniors) were able to revise the diffusion program, which was considered in detail in the class; in pairs to develop additional computational science applications, such as modeling the growth of mushroom fairy rings and modeling animal use of cognitive maps to find food; and to perform well on quiz and test questions on the material.

Moreover, at the end of the term, students had a brief introduction to concurrent processing and parallel algorithms. Consideration of more complex and larger biofilm arrangements, particularly in 3D, illustrated the importance of high performance computing in computational science.

3.5 Class Testing in High Performance Computing Course

The High Performance Computing course (COSC 365) at Wofford was populated by a mixture of Emphasis in Computational Science (ECS) students and computer science majors (five students: one biology/ECS, one chemistry/ECS, one computer science/chemistry/ECS, two computer science; sophomore to senior level). All students had had through Data Structures with programming in Python and C++ and at least one other computer science or computational science course.

With a week of class time devoted to the biofilms module, the class in teams successfully developed MPI/C programs for diffusion using a parallel random number generator on NCSA's Teragrid computer Abe, a Dell Intel 64 Linux Cluster [9]. Later, the students also demonstrated their understanding of the material with their performance on a test and an exam.

In the course, the emphasis is on learning HPC techniques, and diffusion is important in many applications that require high performance computing. Biofilms were particularly interesting to the biology and chemistry students in the class. Moreover, the topic helped computer science majors make the connections of theory to application, which they sometimes overlook, and seemed to provide motivation for all the students. For example, evaluation comments indicated that the module is "very easy to read," and the model "very useful for the real world problems."

3.6 Blue Waters UPEP Internship Involvement

During the summer of 2009, student Shay Ellison had a Blue Waters UPEP Internship to develop parallel versions for two and multiprocessors of the biofilms application and to write a ten-page accompanying tutorial, "Biofilms Parallel Computational Model with MPI and C," which are also available on the NCSI UPEP Curriculum Modules site [2]. Using TeraGrid resources, he investigated parallel random number generation, output of results, and speedup of the program with multiple processors. The experience enhanced his understanding of HPC and undoubtedly is valuable to Shay as he pursues graduate studies in information security at Florida State University.

3.7 Additional Outreach

In an effort to extend the educational benefits of the module, the authors wrote and presented a paper, "Simulating the Formation of Biofilms in an Undergraduate Modeling Course" for the Workshop on Teaching Computational Science at the International Conference on Computational Science (ICCS) in Amsterdam [10].

4. MODULE 2: SOCIAL NETWORKS

4.1 Scientific Question

The module "Getting the 'Edge' on the Next Flu Pandemic: We Should'a 'Node' Better," which has the potential of making similar impacts, covers social networks and individual-based epidemiology simulations. Individual-based (or network-based) epidemiology simulations that track the simulated behavior of individuals in a community provide greater specificity and are easier to verify than cellular automaton simulations [11]. The module discussed the following important metrics for social networks:

- a smallest set of locations (minimum dominating set) that a given proportion of the population visits, which can be helpful in determining sites for fever sensors or in closing of particular public buildings during an epidemic

- the distribution of the number of contacts people have with other people (degree distribution), which can facilitate targeted vaccination of individuals who have many contacts [12]
- the probability that two contacts of a randomly chosen person have contact with one another (clustering coefficient), which is an indication of how rapidly a disease can spread through a community [13]
- the average smallest number of contacts for a disease to spread from one arbitrary individual to another (mean shortest path length), which also indicates the rapidity with which a disease can spread

4.2 Computational Models

The basic data structure for solving this scientific problem is a graph, or a set of nodes with undirected or directed edges connecting some of the points. For a contact network or a social network, the nodes represent people or groups of people, such as members of a household that can become infected, and places, where the disease can spread from an infected person to a susceptible individual (See Figure 2). Each edge represents an association that can lead to transmission of the disease. Thus, besides the biological background, the module covers some of the fundamental concepts in graph theory, such as adjacent nodes, degree, complete graph, and paths. Moreover, students explore some of the characteristics of such social networks and of biological networks in general, such as the following:

- Social networks are scale-free, where most nodes have relatively low degree but a few nodes, called hubs, have high degrees, making such networks particularly vulnerable to attack and failure [12].
- Biological networks exhibit the small world property, where the average length of a path between nodes is small in comparison to the size of the graph, so that these graphs are efficient communicators of information or disease.

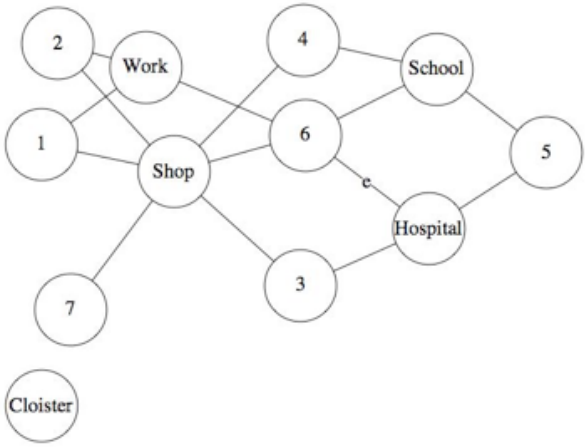

Figure 2. Contact network of households and places

4.3 Algorithms

For implementation of the graph data structure, the module employs a vector to store nodal values, adjacency matrices to represent edge connections and associated values, and connection matrices for existence of associations only. With data of people's activities, we can construct a people-location graph. The module covers the FastGreedy Algorithm to obtain an approximation of a minimum dominating set, or a smallest set of locations that a given proportion of the population visits. Forming a people-people graph of contacts that individuals have with each other by visiting the same locations in a day, we can also compute a distribution of the number of contacts people have with others and the clustering coefficient of each person.

4.4 Assessment of Models

The module and one accompanying *Mathematica* file showed computation of various metrics for 1000 people randomly selected from NDSSL's synthetic data for the population of Portland, Oregon [5]. For example, the degree distribution of this data in Figure 3 approximates the power law, $P(k) = ck^{-r}$, common for scale-free networks. The shape reveals only a few critical nodes have degree 6 or more. The module points out that using further demographic information, public health officials might target such people for immediate vaccination.

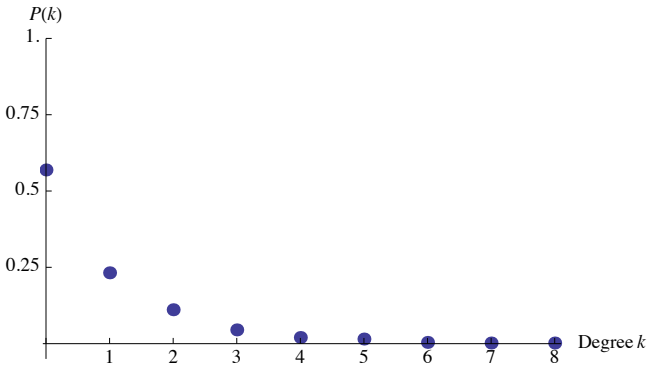

Figure 3. Degree distribution of 1000 randomly selected people

4.5 Need for High Performance Computing

NDSSL's synthetic data for the population of Portland, Oregon involves 1,615,860 people having 8,922,359 activities [5]. Using such data, computational scientists are developing high performance individual-based simulation models to study the nature of epidemics and the impacts of policy decisions on controlling epidemics in urban environments. Because such individual-based simulations incorporate massive amounts of data that require extensive effort to gather and need massive computing power to process, such models provide excellent examples for HPC for the students.

4.6 Class Testing in Mathematical Modeling

During spring 2010, the Mathematical Modeling course at Wofford used this module with *Mathematica* for six class hours, half of which was formal instruction and half hands-on activity in which the students worked through the "base model." Then, deriving their own modifications, students described the physical system and how it was incorporated into the model, solved the model with *Mathematica*, and interpreted the results. Students' learning on this topic was assessed through participation in class activities, journal entries about the reading, and the above assignment.

Class members, typically at the junior-senior level, were two physics, one mathematics/physics, one mathematics/chemistry,

two mathematics majors, and one undecided. The professor, Dr. Anne Catlla, indicated that with background of the module and accompanying files students were able to successfully complete the assignments and that overall there was more variation among individuals than among majors.

A student in the class stated, "I liked the application of social network theory and graph theory to a disease modeling scenario." The professor wrote she "thought that the assignments (both the Quick Review problems and the projects at the end of the module) were excellent;" and in an evaluation she "strongly agreed" that she understood the science applications, mathematics, and algorithms in the module. Although the class did not use HPC, coverage of the material helped to enlighten them on the need for and advantages of using such power for larger data sets and problems.

4.7 Class Testing in High Performance Computing

Over a three-week period at the end of the spring 2010 semester, students in High Performance Computing at Wofford heard from the professor about important graph theory applications and techniques in HPC, studied the epidemics module, presented the material to the rest of the class, and in two teams developed C/MPI programs to calculate various metrics discussed above. The teams implemented their solutions on the Teragrid's NICS Kracken, a 99072-processor Cray XT5 computer [9]. Later, they completed exam questions on the material. All students did well on the module, although the student with the least amount of programming background experienced more difficulty with the HPC aspects of the assignments.

In an evaluation, one student commented, "If detailed enough information was provided and the data set was realistically large, HPC would be invaluable in modeling social networks." Another stated, "I liked being able to see the relation to science in this module." A third student wrote, "I particularly liked the use of a two-dimensional array as a connection graph. We had not really used matrices in such a way before." One of the students was delighted to relate the application to her own research involving the electrical grid in a summer internship at Oak Ridge National Laboratory.

5. CONCLUSION

The limited class sizes preclude statistical analysis at this time, but we are encouraged from the very positive responses from students and professors, alike. We have already incorporated suggestions from participants in all three classes to make the modules more effective. Blue Waters funding in the Undergraduate Petascale Education Program has been invaluable in high performance computing computational science module development, internships, conference participation, teaching, and learning. This project is advancing education personally, locally, nationally, and internationally.

6. REFERENCES

[1] National Computational Science Institute Undergraduate Petascale Education Program (UPEP). http://computationalscience.org/upep

[2] National Computational Science Institute Undergraduate Petascale Education Program (UPEP) Curriculum Modules, UPEP Curriculum Modules site. http://computationalscience.org/upep/curriculum

[3] Wofford College. http://www.wofford.edu

[4] Computational Science - Wofford College. http://www.wofford.edu/computationalscience/

[5] NDSSL (Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University). 2009. "NDSSL Proto-Entities" http://ndssl.vbi.vt.edu/opendata/.

[6] Donlan, R. M. and Costerton, J. W. 2002. Biofilms: survival mechanisms of clinically relevant microorganisms. Clin. Microbiol. Rev. 15 (2) 167-193 ().

[7] Stewart, Philip S. 2003. Guest Commentaries, Diffusion in Biofilms. In *J. Bacteriol*. 185(5): 1485–1491.

[8] Shiflet, A. and Shiflet, G. 2006. *Introduction to Computational Science: Modeling and Simulation for the Sciences*. Princeton University Press, Princeton.

[9] Teragrid. 2010. https://www.teragrid.org/

[10] Shiflet, A. and Shiflet, G. 2010. Simulating the Formation of Biofilms in an Undergraduate Modeling Course ICCS 2010. In *Procedia Computer Science*. Elsevier. 1(1): 895-901.

[11] Bisset, Keith and Marathe, Madhav. 2009. "A Cyber Environment to Support Pandemic Planning and Response." *SciDAC Review* 13: 36-47. http://www.scidacreview.org/0903/html/maranthe.html.

[12] Mason, Oliver and Verwoerd, Mark. 2007. Graph Theory and Networks in Biology. In *IET Systems Biology* 1: 89-119. http://www.hamilton.ie/SystemsBiology/files/2006/graph_theory_and_networks_in_biology.pdf.

[13] Newman, M. E. J., Watts, D. J., and Strogatz, S. H. 2002. Random graph models of social networks *Proceedings of the National Academy of Science* 99 (Suppl 1): 2566-2572. http://www.pnas.org/content/99/suppl.1/2566.full.