Algorithms: From Theory to Application

Theoretically interesting algorithms are increasingly playing a critical role in real-world applications. Shortest path and interior point methods are used in airline scheduling. String matching, traveling salesman and max-flow algorithms are used in computational biology. Algorithms for graph separators, maximal independent set, Delaunay triangulation, and point location are used in finite element codes for designing airplanes, cars, and space structures. Strassen’s matrix multiply is now available as part of many, if not most, matrix libraries. This increasing role is due in part to the need to handle larger data sets, which makes asymptotics more important, and in part to the overall increased sophistication of the techniques used within the application areas requiring a corresponding increased level of sophistication in the algorithms. With the explosion in use and connectivity of computers and data sources, this need for sophisticated algorithms can be expected to increase substantially.

The transition from theory to practice, however, has been slow. It often takes twenty years for an algorithm to make it from theory into an application, and there are still many cases of naive algorithms being used where a better algorithm would surely improve performance and productivity. Furthermore, similar algorithms or techniques are often reinvented many times over in different application areas.

Because of the need to accelerate the transition of ideas from theory to application and back, and the need to better share ideas and code for basic algorithms and data structures among diverse applications, we believe it is critical to support broad-based research on algorithms from theory to applications. This research would span several of the application areas outlined in the ITR program solicitation, including information management, software, advanced computational science, and scalable information infrastructure.

The intellectual core of this proposal is to study algorithms in an integrated manner from pure theory to applications. The primary goals are (1) to improve connections between application areas by sharing algorithms of common interest, (2) to expose and tackle fundamental issues common to disparate areas, and (3) to develop fundamentally new methods for solving critical problems via the connections being made. Our approach will be to form a tight research group that, in addition to solidifying interactions among its members, will build close collaborations with researchers in various application areas. This interaction will be achieved through joint papers, workshops, common weekly meetings, shared students, a shared repository of algorithms, and the development of new course material.

Background

The proposed research is motivated in part by a course we have developed titled “Algorithms in the Real World” (http://www.cs.cmu.edu/~guyb/realworld.html), and a workshop we are organizing on the same topic (http://www.cs.cmu.edu/~guyb/rwc/). The course covers algorithms for compression, cryptography, optimization, triangulation, VLSI, computational biology, and indexing. In developing the material for the course we noticed that theoretically interesting algorithms are being used in practice much more than most people expect. As an example, we have found that Splay Trees (a data structure invented by PI Danny Sleator) are used heavily in Windows/NT, Linux, the gcc compiler, AT&T’s and Fore system’s network.
routers, the Lotus word processor, the Unix malloc routine, the SQUIB web proxy cache, as well as many other applications. We also noticed, however, that in recent years research on algorithms has often taken on a life of its own within a variety of application areas. Although this reflects well on the importance of non-trivial algorithms in practice, it has meant that many techniques are reinvented many times over. For example, singular-value decomposition is used in a similar way in Computer Vision, Information Retrieval, Data Compression, and Computational Biology. Most of the work in these areas has proceeded independently without reference to the other areas.

**Research Directions**

To limit the scope of the proposed work, we plan to emphasize certain problem domains based on the backgrounds of the PIs and the potential for collaboration with industry and other research groups within Carnegie Mellon University. Specific problem domains we plan to study include computational geometry, data mining, robotics, graphics, security, computational biology, and information retrieval. Three specific topics are described in more detail here.

**Algorithms for Indexing and Searching:** Efficient indexing and searching of the exponentially growing archives of on-line information will require increasingly sophisticated algorithms. Information retrieval and management involves almost every aspect of algorithms, including data structures, random walks, graph algorithms, compression, cryptography, load-balancing, and many others. A survey of some of the important and technically interesting research challenges in this area was presented at a recent tutorial at FOCS (Broder and Henzinger, 1998).

Certainly many of the algorithmic aspects involved in classical information retrieval for Boolean queries and the vector space model (Salton, Wong and Yang, 1975) are by now well understood and engineered in practical systems (Witten et al., 1999). However, even in this well studied area, recent research on algorithms may be able to yield improvements. For example, binary operations on ordered sets (union, intersection, etc.) using random balanced binary trees and parallel algorithms can result in significant speed-ups over more standard techniques (Blloch and Reid-Miller, 1999). As another example, recent research on document ranking has made use of the singular-value decomposition for matrices (Berry et al., 1995; Kleinberg, 1998; Bharat and Henzinger, 1998), and these methods have much in common with graph separator algorithms (Guattery and Miller, 1995), Monte-Carlo algorithms for low-rank approximations (Frieze et al., 1998) and other topics of active interest in the algorithms community.

But information retrieval and management is rapidly evolving in ways that will raise many new algorithmic issues. For example, new probabilistic approaches to information retrieval are now emerging (Ponte and Croft, 1998; Berger and Lafferty, 1999; Miller et al., 1999) and the algorithmic aspects of scaling these approaches to very large data collections are virtually unexplored. Moreover, the way in which a user is able to search for information is evolving. Current research efforts are going beyond simple Boolean queries to more sophis-
ticated question answering systems (Singhal, 1999) and context-sensitive queries. Efficiently handling such searching paradigms will require new ways of indexing and retrieving data. One of the goals of this research project will be to open communication between researchers in the information retrieval and algorithms communities, so that good solutions to these emerging problems can be found quickly, without “reinventing the wheel” on many of the more basic algorithmic issues.

**Human Oriented ID Project:** The Human Oriented ID Project (HumanOID) is a fundamentally new approach, arising from connections between disparate areas of Machine Learning and Cryptography, to solve a pressing need: secure, easy to use authentication. Authentication is the process by which a user must convince a computer — usually by presenting the proper password — that the user is who he says he is. Password-based authentication is ubiquitous in a broad range of applications.

Unfortunately, passwords, including fingerprints and retinal scans, have the potential to be recorded by an eavesdropper. One way to make authentication secure against eavesdropping is to use a challenge-response system. In challenge-response, a computer presents one of a large set of possible challenges to a user, who then provides a correct response. If properly designed, an eavesdropper who observes a large number of challenge-response pairs will not be able to impersonate the user on a new challenge.

Secure challenge-response protocols have been around for over a decade, based on public-key cryptography or zero-knowledge proofs. However, these require the user to perform computations that themselves need a separate computing device or smart-card. The goal of the HumanOID project, headed by PIs M. Blum and A. Blum, is to develop a challenge-response protocol that a user can perform easily in his head, without need for any added computing device. In addition to the advantage over passwords in terms of security against eavesdropping, there is also potential to actually reduce cognitive load on users who currently need to remember multiple unrelated passwords for security purposes. (For a description of the growing problem caused by this cognitive load, see (Adams and Sasse, 1999).)

The HumanOID project is based on a merging of Cryptography and Machine Learning. The connection between machine learning and authentication is this: In a challenge-response system, an eavesdropper is trying to learn how to create a valid response to a new challenge by observing many challenge-response pairs. To prevent the eavesdropper from learning, a good challenge-response authentication protocol should therefore be based on a “simple but hard” learning problem. Formal connections between these areas have been made already in (A. Blum, Furst, Kearns, and Lipton, 1994). This project intends to turn some of these high-level theoretical connections into a real system usable by real people.

**Algorithms for Triangulated Meshing:** Triangulated meshes are becoming increasingly important in a wide variety of application areas, including graphics, geographic information systems, computer vision, data compression, and engineering and scientific simulations. Triangulated meshes lead to many interesting theoretical questions, and make use of a large number of basic algorithms as subcomponents. For example maximum-flow and maximal-independent sets on graphs have both been used in mesh coarsening algorithms (Miller, Talmor, and Teng, 1997; Adams, 1998). Although the basic techniques such as Delaunay
triangulation are shared among all the application areas, much of the code and many of the more advanced techniques are developed independently in each of the application areas.

The PIs plan to collaborate with three researchers/projects at Carnegie Mellon to see if common tools can be shared among the groups. The researchers are Omar Ghattas of the Quake project, who is making use of triangulated meshes to simulate ground motion in earthquakes, Paul Heckbert of the graphics group, who is using triangulated meshes for surface approximation and multi-resolution rendering, and Takeo Kanade of the vision group, who is using triangulated meshes for 3d surface reconstruction from stereo vision. We have already been working with Omar Ghattas on making use of their parallel Delaunay code (Blelloch, Hardwick, Miller and Talmor 1999) for dynamically meshing highly deforming structures. Furthermore they have been using one of Heckbert’s surface simplification algorithms to study methods for persistent triangulations (Blelloch et al., 1999).

The PIs and the Carnegie Mellon environment

The PIs have a broad background in algorithms, both theoretical and applied as described below. Furthermore the School of Computer Science at Carnegie Mellon is an ideal environment for the proposed research because of the extensive ongoing activity in the applications areas to be studied, including machine learning, robotics, and information retrieval.

Guy Blelloch: known for his development of the NESL programming language, as well as fast parallel algorithms for sorting, Delaunay triangulation, list-ranking, and graph connectivity.

Avrim Blum: developer of the Graphplan planning algorithm, and known for his research in algorithms for machine learning, data mining, and combinatorial optimization.

Manuel Blum: winner of the ACM Turing Award for his work in the “foundation of computational complexity theory and its applications to cryptography and program checking.”

John Lafferty: known for his work in language modeling and information retrieval, and co-developer (along with PI Sleator) of the Link Grammar natural-language parser.

Danny Sleator: developer of the Splay Tree data structure used in applications from network routers to the Lotus word processor, who more recently has been applying algorithm ideas to Computer Music and Natural Language.

The PIs will also be collaborating with colleagues including Gary Miller (algorithms and scientific computing), Andrew Moore (datamining), R. Ravi (optimization and computational biology), and Jeannette Wing (security and formal methods).

Research Results and Impact

The results and expected impact of the proposed research will include the following.
• We will develop new algorithms and techniques, with an emphasis on techniques that span several application areas. We also plan to implement the algorithms we develop and make them available as libraries.
• We expect to explore new theoretical problems by considering common ideas that appear in different application areas.
• We will generate on-line course material that connects the study of algorithms to the uses of algorithms in disparate applications. This material will be available via the web to undergraduate and research institutions across the country.
• We expect that our research will lead to better sharing of algorithms and techniques among different application areas.

Integration of Research and Education

We plan to integrate the proposed work with educational activities in several ways. First, we plan to continue the “Algorithms in the Real World” course, and to develop a textbook and other materials based on it. The proposed research will help guide the course, and similarly, course projects are good opportunities to have students implement new algorithms and try out new ideas that will contribute to the research. Second, we plan to run workshops on applied algorithms; we are running one such workshop this May. These workshops will bring together researchers in academia and industry who are developing and using algorithms in their work. Third, we plan to initiate a summer institute for college undergraduates of computer science. It would introduce junior-senior level college undergraduates from around the world to great ideas in computer science, and to research in algorithms in particular.

We also believe that our project has a number of features that make it ideally suited to help increase the participation of underrepresented groups in computer science: The area itself, “algorithms in the real world,” has broad appeal and draws on a broad range of interests and expertise. The field is new and relatively accessible. Our university is strongly committed to increasing diversity. In computer science we have already implemented several successful programs. For a number of years, our colleagues (under the direction of Allan Fisher) have run summer programs for high school teachers of advanced placement computer science which, in addition to providing technical expertise, have stressed ways of increasing the participation of young women. As a consequence, this year over 37% of the first year class in the Computer Science Department is female. Carnegie Mellon President Cohon has provided a large grant for projects (under the direction of Lenore Blum) aimed at ensuring the ongoing success of our female students. Building on these experiences, we are looking at ways of increasing the participation of other underrepresented groups. One way will be to actively recruit such participation in our upcoming conference “Algorithms in the Real World” and in the proposed summer program in this area for undergraduates.

We anticipate that the web site to be developed in conjunction with our project will serve as a resource for many undergraduate programs across the country. By providing exciting new “real world” material that can be readily incorporated into various courses we hope to be able to impact programs and students beyond our campus, including those at minority institutions.