Preface

With the unprecedented rate at which data is being collected today in almost all fields of human endeavor, there is an emerging economic and scientific need to extract useful information from it. For example, many companies already have data-warehouses in the terabyte range (e.g., FedEx, Walmart). The World Wide Web has an estimated 800 million web-pages. Similarly, scientific data is reaching gigantic proportions (e.g., NASA space missions, Human Genome Project). High-performance, scalable, parallel, and distributed computing is crucial for ensuring system scalability and interactivity as datasets continue to grow in size and complexity.

To address this need we organized the workshop on Large-Scale Parallel KDD Systems, which was held in conjunction with the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, on August 15th, 1999, San Diego, California. The goal of this workshop was to bring researchers and practitioners together in a setting where they could discuss the design, implementation, and deployment of large-scale parallel knowledge discovery (PKD) systems, which can manipulate data taken from very large enterprise or scientific databases, regardless of whether the data is located centrally or is globally distributed. Relevant topics identified for the workshop included:

- How to develop a rapid-response, scalable, and parallel knowledge discovery system that supports global organizations with terabytes of data.
- How to address some of the challenges facing current state-of-the-art data mining tools. These challenges include relieving the user from time and volume constrained tool-sets, evolving knowledge stores with new knowledge effectively, acquiring data elements from heterogeneous sources such as the Web or other repositories, and enhancing the PKD process by incrementally updating the knowledge stores.
- How to leverage high performance parallel and distributed techniques in all the phases of KDD, such as initial data selection, cleaning and preprocessing, transformation, data-mining task and algorithm selection and its application, pattern evaluation, management of discovered knowledge, and providing tight coupling between the mining engine and database/file server.
- How to facilitate user interaction and usability, allowing the representation of domain knowledge, and to maximize understanding during and after the process. That is, how to build an adaptable knowledge engine which supports business decisions, product creation and evolution, and leverages information into usable or actionable knowledge.

This book contains the revised versions of the workshop papers and it also includes several invited chapters, to bring the readers up-to-date on the recent developments in this field. This book thus represents the state-of-the-art in parallel and distributed data mining methods. It should be useful for both researchers

and practitioners interested in the design, implementation, and deployment of large-scale, parallel knowledge discovery systems.

December 1999

Mohammed J. Zaki
Ching-Tien Ho

Workshop Chairs

Workshop Chair: Mohammed J. Zaki (Rensselaer Polytechnic Institute, USA)
Workshop Co-Chair: Ching-Tien Ho (IBM Almaden Research Center, USA)

Program Committee

David Cheung (University of Hong Kong, Hong Kong)
Alok Choudhary (Northwestern University, USA)
Alex A. Freitas (Pontifical Catholic University of Parana, Brazil)
Robert Grossman (University of Illinois-Chicago, USA)
Yike Guo (Imperial College, UK)
Hillol Kargupta (Washington State University, USA)
Masaru Kitsuregawa (University of Tokyo, Japan)
Vipin Kumar (University of Minnesota, USA)
Reagan Moore (San Diego Supercomputer Center, USA)
Ron Musick (Lawrence Livermore National Lab, USA)
Srini Parthasarathy (University of Rochester, USA)
Sanjay Ranka (University of Florida, USA)
Arno Siebes (Centrum Wiskunde Informatica, Netherlands)
David Skillicorn (Queens University, Canada)
Paul Stolorz (Jet Propulsion Lab, USA)
Graham Williams (Cooperative Research Center for Advanced Computational Systems, Australia)

Acknowledgements

We would like to thank all the invited speakers, authors, and participants for contributing to the success of the workshop. Special thanks are due to the program committee for their support and help in reviewing the submissions.