DIMACS Center
Rutgers University

Workshops in Data Analysis and Mining

Final Report

September 2005

Ia. Participants in the program

Organizer:

Fred Roberts, DIMACS

Workshop on Data Quality, Data Cleaning and Treatment of Noisy Data
Dates: November 3 - 4, 2003

Organizer:

Parni Dasu, AT&T

Workshop on Discrete Metric Spaces and their Algorithmic Applications
Dates: August 20 - 23, 2003

Organizers:

Moses Charikar, Princeton University
Piotr Indyk, MIT
Nati Linial, Hebrew University
Jiri Matousek, Charles University
Yuri Rabinovich, University of Haifa
Gideon Schechtman, Weizmann Institute

Working Group Meeting on New Algorithms for Inferring Molecular Structure from Distance Restraints
Dates: January 12 - 16, 2004

Organizer:

Michael W. Trosset, College of William & Mary

Working Group Meeting on The Burrows - Wheeler Transform: Ten Years Later
Dates: August 19 - 20, 2004

Organizers:

Paolo Ferragina, University of Pisa
Giovanni Manzini, University of Piemonte Orientale
S. Muthukrishnan, Rutgers University

Ib. Participating Organizations

Telcordia Technologies: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

AT&T Labs - Research: Collaborative Research

Partner organization of DIMACS. Individuals from the organization participated in the program planning, research and Working Group leadership.

NEC Laboratories America: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Lucent Technologies, Bell Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Princeton University: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research. Co-sponsored the Workshop on Discrete Metric Spaces and their Algorithmic Applications.

Avaya Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

HP Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

IBM Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Microsoft Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Institute for Advanced Study
Co-sponsored Workshop on Discrete Metric Spaces and their Algorithmic Applications

1c. Other Collaborators

The project involved scientists from numerous institutions in numerous countries. The resulting collaborations also involved individuals from many institutions in many countries.

II. Project Activities

Workshop on Data Quality, Data Cleaning and Treatment of Noisy Data
Dates: November 3 - 4, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Parni Dasu, AT&T
Attendance: 46

The word “data” has taken on a broad meaning in the last five years. It is no longer a set of numbers or even text. New data paradigms include data streams characterized by a high rate of accumulation, web scraped documents and tables, web server logs, images, audio and video, to name a few. Well-known challenges of heterogeneity and scale continue to grow as data are integrated from disparate sources and become more complex in size and content.

While new paradigms have enriched data, the quality of data has declined considerably. In earlier times, data were collected as a part of pre-designed experiments where data collection could be monitored to enforce data quality standards. The data sets themselves were small enough that even if data collection was unsupervised, the data could be quickly scrubbed through highly manual methods. Today, neither monitoring of data collection nor manual scrubbing of data is feasible due to the sheer size and complexity of the data.

An additional challenge in addressing data quality is the domain dependence of problems and solutions. Metadata and domain expertise have to be discovered and incorporated into the solutions, entailing an extensive interaction with widely scattered experts. This particular aspect of data quality makes it difficult to find general one-size-fits-all solutions. However, the process of discovering metadata and domain expertise can be automated through the development of appropriate tools and techniques such as data browsing and exploration, knowledge representation and rule based programming.

Many disciplines have taken piecemeal approaches to data quality. The areas of process management statistics, data mining database research and metadata coding have all developed their own ad hoc approaches to solve different pieces of the data quality puzzle. These include statistical techniques for process monitoring, treatment of incomplete data and outliers, techniques for monitoring and auditing data delivery processes, database research for integration, discovery of functional dependencies and join paths, and languages for data exchange and metadata representation.

We need an integrated end-to-end approach within a common framework, where the various disciplines can complement and leverage each other's strengths. In this workshop, our broad objective was to bring together experts from different research disciplines to initiate a comprehensive technical discussion on data quality, data cleaning and treatment of noisy data. Specifically,

* To provide an overview of the existing research in data quality
* To present data quality as a continuous, end-to-end concept
* To discuss and update the definition of data quality, to develop metrics for measuring data quality
* To emphasize data exploration, data browsing and data profiling for validating schema specific constraints and identifying aberrations
* To focus on disciplines such as knowledge representation and rule based programming for capturing and validating domain specific constraints
* To highlight applications, case studies
* To present research tools and techniques
* To identify research problems in data quality and data cleaning

The format of the workshop was a combination of invited talks, contributed papers and posters.

Workshop on Discrete Metric Spaces and their Algorithmic Applications

Dates: August 20 - 23, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Moses Charikar, Princeton University; Piotr Indyk, MIT; Nati Linial, Hebrew University; Jiri Matousek, Charles University; Yuri Rabinovich, University of Haifa; Gideon Schechtman, Weizmann Institute
Attendance: 97

Co-sponsored by the Institute for Advanced Study and Princeton University.

The study of finite metric spaces and metric embeddings has its origins in Banach space theory and functional analysis. In recent years, this area has emerged as a new and influential branch of discrete mathematics, with deep and surprising applications in Computer Science. Theoretical computer scientists are now familiar with the work of Bourgain, who showed that any metric on n points can be embedded with O(log n) distortion in Euclidean space, as well as the fundamental dimension reduction result of Johnson and Lindenstrauss who showed that any n points in Euclidean space can be embedded in $O(\log(n)/ \varepsilon^2)$ dimensions with a $(1 + \varepsilon)$- distortion in distances. Following the work of Linial, London and Rabinovich, Bourgain's result has found novel applications to approximation algorithms. The Johnson Lindenstrauss lemma has been used as a basic tool for dealing with high dimensional data. It has also been used extensively in the design of algorithms and data structures for various problems in high dimensional computational geometry.

There has been a lot of interest and activity in this area in the past few years. New insights into metric embeddings have been developed as well as new applications and connections discovered to approximation algorithms, algorithms for large data sets, and algorithms for high dimensional computational geometry. Recently, there have been several surveys on the subject. Piotr Indyk gave a tutorial at FOCS 2001 on algorithmic applications of low distortion embeddings. Jiri Matousek wrote a chapter on mathematical aspects of embeddings in his recent book on discrete geometry. Nati Linial gave a survey on finite metric spaces and their connection to combinatorics, geometry and algorithms at the International Congress of Mathematicians in 2002. Also, Piotr Indyk and Jiri Matousek recently wrote a chapter on low distortion embeddings of finite metric spaces for the 2004 Handbook of Discrete and Computational Geometry.

The questions in this field are a beautiful blend of combinatorics and geometry with several connections to algorithm design and analysis. Here are some of the topics that the workshop covered:

- Embeddings in normed spaces and connections to Banach space theory.
- Embeddings of planar graphs, low distortion embeddings into trees and their applications to approximation algorithms.
- Metric Ramsey type theorems.
- Low dimensional embeddings and connections to algorithmic techniques for large data sets.
- Embeddings of special metrics (e.g. edit distance, Frechet distance, etc.) and applications to high dimensional computational geometry.
- Graph representations, distance labelings, graph spanners, etc.

Working Group Meeting on New Algorithms for Inferring Molecular Structure from Distance Restraints
Dates: January 12 - 16, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Michael W. Trosset, College of William & Mary
Attendance: 7

An important problem in structural molecular biology is the problem of determining 3-dimensional molecular structure using NMR spectroscopy. One critical step in inferring molecular structure from NMR data involves computing a 3-dimensional configuration of points that is consistent with a specified set of lower and upper bounds on the interpoint distances. This step can be formulated, in various ways, as a numerical optimization problem. This research week allowed several researchers who were developing new algorithms for such problems to interact and collaborate, and has led to several co-authored papers.

Working Group Meeting on The Burrows - Wheeler Transform: Ten Years Later

Dates: August 19 - 20, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Paolo Ferragina, University of Pisa; Giovanni Manzini, University of Piemonte Orientale; S. Muthukrishnan, Rutgers University
Attendance: 34

On May 10, 1994 Michael Burrows and David Wheeler published the Technical Report "A block sorting lossless data compression algorithm" describing a new data compression algorithm based on a reversible transformation of the input. This transformation had a profound impact in several algorithmic fields and is today universally known as the BWT: the Burrows-Wheeler Transform.

The BWT changed the approach to data compression in a twofold way: it highlighted the role of input transformations to prepare data for achieving better compression, and showed that a pipeline of simple encoders may be more effective than a one-shot compression pass. The BWT is at the heart of the algorithm bzip2, which has become the standard tool for lossless compression, thanks to its 20% compression on average against the 33% for gzip. Ten years after its discovery, new algorithms and theoretical studies deploying the BWT continue to flourish. The BWT has also revolutionized the field of indexing data structures: using the BWT it is possible to build the so called "compressed indexes," a new family of data structures that support powerful substring searches and occupy roughly the same space required by the best compressors. These results have disproved the belief that an efficient full-text index must use space superlinear in the indexed string length (cfr. suffix tree and array).

Recent theoretical results, for example on combinatorics on words, have shown that our understanding of the properties of the BWT and its potential applications is still incomplete. This workshop promoted the further study of the theory and applications of this fascinating mathematical tool.

III. Project Findings

A Sublinear Algorithm of Sparse Fourier Transform for Nonequispaced Data

Jing Zou (Princeton University) was a participant in the Workshop on Discrete Metric Spaces and their Algorithmic Applications, which influenced his work on the problem in which the recovery of a discrete time signal S of length N is sought when only L signal values are known. In general, this is an insoluble problem. Zou considered it under the additional assumption that the signal has a sparse Fourier transform. Using results in the papers Mansour, "Randomized interpolation and approximation of sparse polynomails," SICOMP, 24(2), 1995, Akavia, Goldwasser and S. Safra, "Proving hard-core predicates by list decoding", 146--157, FOCS 2003, and Gilbert, Guha, Indyk, Muthukrishnan and Strauss, "Near optimal sparse Fourier representation via sampling," STOC 2002, Zou developed a sublinear randomized algorithm to recover a B-sparse Fourier representation of a signal S from incomplete data. The algorithm also extends to the case where the Fourier transform is not B-sparse, where the aim is to find a near-optimal B-term Fourier representation. A typical application of this work is the observation of non-equi-spaced data. For a signal with evenly spaced data, the famous Fast Fourier Transform (FFT) computes all the Fourier coefficients in time O(N logN). However, the requirement of equally distributed data by FFT raises challenges for many important applications. For instance, because of the occurrence of instrumental drop-outs, the data may be available only on a set of non-consecutive integers. Another example occurs in astronomy, where the observers cannot completely control the availability of observational data: a telescope can only see the universe on nights when skies are not cloudy. In fact, computing the Fourier representation from irregularly spaced data has wide applications in processing astrophysical and seismic data, the spectral method on adaptive grids, the tracking of Lagrangian particles, and the implementation of semi-Lagrangian methods. In many of these applications, a few large Fourier coefficients already capture the major time invariant wave-like information of the signal, and very small Fourier coefficients can be ignored. To find a small set of the largest Fourier coefficients and hence a (near) optimal B-sparse Fourier representation of a signal that describes most of the signal characteristics is a fundamental task in applied Fourier Analysis. Zou's algorithm reconstructs a near-optimal B-term representation. Numerical experiments support the advantage in speed of this algorithm over other methods for sparse signals: it already outperforms Inverse Nonequispaced Fast Fourier Transform (INFFT) for large but realistic size N and works well even in the situation of a large percentage of missing data and in the presence of noise.

Theoretical and Experimental Analysis of a Randomized Algorithm for Sparse Fourier Transform Analysis

In many situations, a few large Fourier coefficients already capture the major time-invariant wave-like information of a signal and very small Fourier coefficients can thus be discarded. The problem of finding the (hopefully few) largest Fourier coefficients of a signal that describe most of the signal trends is a fundamental task in Fourier Analysis. Techniques to solve this problem are very useful in data compression, feature extraction, finding approximating periods and other data mining tasks, as well as in situations where multiple scales exist in the domain (as in e.g. materials science), and the solutions have sparse modes in the frequency domain. Jing Zou (Princeton University), Anna Gilbert (AT&T Labs-Research), Martin Strauss (AT&T Labs-Research), and Ingrid Daubechies (Princeton University) analyzed a sublinear RAℓSFA (Randomized Algorithm for Sparse Fourier Analysis) that finds a near-optimal B-term Sparse Representation R for a given discrete signal S of length N, in time and space poly(B, log(N)). Its time cost poly(log(N)) should be compared with the superlinear (N logN) time requirement of the Fast Fourier Transform (FFT). A straightforward implementation of the RAℓSFA turns out to be very slow in practice. Zou, Gilbert, Strauss, and Daubechies' main result is a greatly improved and practical RAℓSFA. They introduce several new ideas and techniques that speed up the algorithm. Note that this built on the previous work of Mansour, Akavia, Goldwasser, S. Safra, Gilbert, Guha, Indyk, Muthukrishnan and Strauss, as mentioned above.

Succint Representations of Sequences

Recently there has been an increased interest in succinct data structures. The aim is to represent data using as little space as possible, yet efficiently answer queries on the represented data. Several results exist on the representation of sequences, trees, graphs, permutations, etc. A heavily studied case is that of binary sequences with rank and select queries. The case of general sequences has received less attention. Paolo Ferragina (Università di Pisa), Giovanni Manzini (Università del Piemonte Orientale), Veli Makinen (University of Helsinki) and Gonzalo Navarro (University of Chile) generalized the results on binary sequences to sequences of symbols in the range [1, r] for small r. The main challenge was to generate short descriptions of pieces of the sequence, which can be computed in constant time and can be used to index into tables with partial precomputed queries. This is more complex than for binary sequences. This result in turn allowed Ferragina, Manzini, Makinen, and Navarro to generalize wavelet trees, improving time to answer queries from logarithmic to constant. Finally, they showed how this result can be used for text indexing. A full-text self-index is a succinct data structure that represents a text string while supporting not only access to any character but also the efficient search for an arbitrary pattern as a substring of the indexed text. Using their new sequence representation they achieve the same space usage as the best previous method and an improvement on time complexities. Their result is the fastest among those full text indexes that use minimum space.

Structuring Labeled Trees for Optimal Succinctness, and Beyond

Consider an ordered, static tree T on t nodes where each node has a label from alphabet set A. Tree T may be of arbitrary degree and of arbitrary shape. Suppose we wish to support basic navigational operations such as find the parent of node u, the i-th child of u, and any child of u with label a. Over fifteen years ago, Jacobson observed that pointer-based tree representations are wasteful in space and introduced the notion of succinct data structures. He studied the special case of unlabeled trees and presented a succinct data structure of *2t+o(t)* bits supporting navigational operations in *O(1)* time. The space used is asymptotically optimal with the information-theoretic lower bound averaged over all trees. This led to a slew of results on succinct data structures for arrays, trees, strings and multisets. Still, for the fundamental problem of structuring labeled trees succinctly, few results, if any, exist even though labeled trees arise frequently in practice, e.g. in the data as in markup text (XML) or in augmented data structures. Paolo Ferragina (Università di Pisa), Fabrizio Luccio (Università di Pisa), Giovanni Manzini (Università del Piemonte Orientale), and S. Muthukrishnan (Rutgers University) presented a novel approach to the problem of succinct manipulation of labeled trees by designing what they call the XBW transform of the tree, in the spirit of the well-known Burrows-Wheeler transform for strings. XBW transform uses path-sorting and grouping to linearize the labeled tree *T* into two coordinated arrays, one capturing the structure and the other the labels. Using the properties of the XBW transform, they (i) derive the first-known (near-optimal) results for succinct representation of labeled trees with *O(1)* time for navigation operations, (ii) optimally support the powerful subpath search operation for the first time, and (iii) introduce a notion of tree entropy and present linear time algorithms for compressing a given labeled tree up to its entropy beyond the information-theoretic lower bound averaged over all tree inputs. Their XBW transform is simple and likely to spur new results in the theory of tree compression and indexing, and may have some practical impact in XML data processing.

Further Work on The Burrows - Wheeler Transform
Peter Fenwick's participation in The Working Group on The Burrows - Wheeler Transform: Ten Years Later led to further work in this area. Fenwick used a BW transform to derive all contexts of some fixed order, say 4, and used these to drive an escape-free PPM compressor/decompressor. He "broke open" the reverse BW transform to generate partial contexts during both compression and decompression, so aiding the operation of the coder and decoder. He developed analogies between the Fourier Transform and its associated spectra, autocorrelations, etc, on the one hand and the Burrows Wheeler transform and PPM on the other. Invoking ideas of symmetry led to a significant new understanding of the Burrows Wheeler operation. Fenwick has also explored the relation between data compression and error correction codes. Compression can be thought of as a very noisy channel (encoding a single symbol into a range of possible symbols) and then generation of the codes to select the correct symbol from the "noise" codes; the encoder and decoder track each other and the "corrections" are what the compressor emits. This idea leads to two other ideas relating to the possible relationship between error correction and compression. Can we omit some symbols (giving an erasure code) and then use a Viterbi-like trellis to recover feasible decodings? The relatively new error-correcting "turbo-codes" use two interacting but complementary error-correcting codes to give superlative performance. The methods developed show promise in answering the question: Can something similar be done for compression?

Realizability of Graphs
A finite graph, without loops or multiple edges, is d-realizable if it can be drawn in d-dimensional space in such a way that every edge has the same length. The general problem is to characterize those graphs that are d-realizable and to find an algorithm that determines whether a graph is d-realizable in polynomial time. The most interesting case, and the one with the most applications, is d=3. A minor H of a graph G is obtained by successively contracting and/or deleting edges of G. The graph K_n is the complete graph on n vertices; V_8 is the graph obtained by joining opposite vertices of a cycle of length 8, and C_5 x C_2 is two copies of a cycle of length 5, with corresponding vertices joined by edges. The class of 3-realizable graphs is minor-closed. It was known that the list of minimal forbidden minors included K_5, the octahedron, and possibly the graphs V_8 and C_5 x C_2. Robert Connelly (Cornell University) and Maria Sloughter collaborated in showing that in fact V_8 and C_5 x C_2 are 3-realizable.

Nonlinear Type and Cotype
One of the deepest aspects of the local theory of Banach spaces is the theory of type and cotype. Various problems that have to do with the development of nonlinear analogs of these concepts have remained

open for many years. Hilbert space is characterized among all normed spaces, by the parallelogram identity. The power of this identity has led researchers to study several weakenings of it, including type and cotype. These parameters control numerous geometric properties of a normed space. The notion of Markov type, defined for a metric space, is useful in various embedding problems. When restricted to normed spaces, the relation between Markov type and linear type is unclear. Assaf Naor (Microsoft Research), Yuval Peres (UC Berkeley), Oded Schramm (Microsoft Research), and Scott Sheffield (UC Berkeley) have shown that L_p, $p>2$, has Markov type 2 and also that trees have Markov type 2. Assaf Naor (Microsoft Research) and Manor Mendel (Hebrew University) have made progress in finding a nonlinear version of cotype. They defined a metric cotype and proved several satisfactory properties for it.

IV. Project Training/Development

Meeyoung Cha, a student from KAIST, South Korea, was a visitor at DIMACS, August 2-27, 2004 supported by this project. Cha worked on the problem of guaranteeing fast packet transmission in an IP network using path diversity. Real-time applications such as Voice over IP (VoIP) need fast and reliable packet transmission even under link and router failures. Using path diversity in the network, such interactive applications can gain end-to-end Quality of Service (QoS). Cha worked with Aman Shaikh (AT&T Labs – Research) while at DIMACS and plans on continuing this collaboration. Cha's papers are listed in the section on Papers/Books/Internet.

Funds were also available at each workshop for support of non-local students.

V. Outreach Activities

Project visitors, graduate students, and senior faculty interacted with 2- and 4-year college faculty in the DIMACS "Reconnect" program, with high school teachers in the DIMACS Connect Institute and the DIMACS Bio-Math Connect Institute, and with undergraduates in the DIMACS Research Experiences for Undergraduates.

VI. Papers/Books/Internet

Book(s) of other one-time publication(s):

A special issue of the journal **Theoretical Computer Science** devoted to the Burrows-Wheeler Transform and its applications edited by P. Ferragina, S. Muthukrishnan, and G. Manzini will be published in 2006.

Journal and Proceedings Articles from all Three Working Groups

M. Cha, S. Moon, C. Park, and A. Shaikh, "Placing relay nodes for intra-domain path diversity," KAIST CS-TR 2005-214.

M. Cha, S. Moon, C. Park, and A. Shaikh, "Placing relay nodes for intra-domain path diversity," submitted.

M. Charikar and A. Karagiozova, "A tight threshold for metric Ramsey phenomena," in *Proc. 16th ACM-SIAM Symp. Discr. Algo.*, 2005, to appear.

G. Cormode and S. Muthukrishnan, “Substring compression problems,” in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms,* 2005.

P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro, “An alphabet friendly FM-index,” in *Proc. 11th Symposium on String Processing and Information Retrieval (SPIRE '04),* Padova, Italy, Lecture Notes in Computer Science, Springer Verlag, **3246** (2004), 150-160.

P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro, “Succinct representations of sequences,” Technical Report DCC-2004-5, University of Chile, August, 2004.

P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro, “Succinct representations of sequences and full-text indexes,” submitted.

P. Ferragina, F. Luccio, G. Manzini, and S. Muthukrishnan, “Structuring labeled trees for optimal succinctness, and beyond,” in *Proc. 46th IEEE Symposium on Foundations of Computer Science (FOCS '05).* Pittsburgh (PA), 2005, to appear.

I.G. Grooms, R.M. Lewis, and M.W. Trosset, “A squared dissimilarity parameterized approach to the embedding problem in distance geometry,” in preparation.

R.M. Lewis and M.W. Trosset, “Sensitivity analysis of the strain criterion for multidimensional scaling,” *Computational Statistics and Data Analysis*, special issue devoted to Matrix Computations and Statistics II, **50** (2006), 135-153, to appear.

R.M. Lewis and M.W. Trosset, “Extensions of classical multidimensional scaling: Variable alternation and nonconvex duality,” *SIAM Journal on Optimization*, March 2005, submitted.

M. Mendel and A. Naor, “Metric cotype,” submitted. Extended abstract, in *17th Annual ACM-SIAM Symp. on Discrete Algorithms*, 2006, to appear.

A. Noar, Y. Peres, O. Schramm and S. Sheffield, “Markov chains in smooth Banach spaced and Gromov hyperbolic metric spaces,” preprint, 2004.

M. Sloughter, “Realizability of graphs in three dimensions,” Cornell University, preprint, 2004.

J. Zou “Theoretical and experimental analysis of a randomized algorithm for sparse fourier transform analysis,” *Journal of Computational Physics*, submitted.

J. Zou, “A sublinear algorithm of sparse fourier transform for nonequispaced data”, *Applied and Computational Harmonic Analysis*, submitted.

Talks/Posters

M. Cha, S. Moon, C-D. Park, and A. Shaikh, “Positioning relay nodes in ISP networks,” IEEE INFOCOM (Poster Session), Miami, USA, March 2005.

VII. Other Products

The main web page for the Special Focus on Data Analysis and Mining is:
http://dimacs.rutgers.edu/SpecialYears/2001_Data/.

The web pages for the workshops and working groups in this project are:

Workshop on Data Quality, Data Cleaning and Treatment of Noisy Data
http://dimacs.rutgers.edu/Workshops/DataCleaning/

Workshop on Discrete Metric Spaces and their Algorithmic Applications
http://dimacs.rutgers.edu/Workshops/MetricSpaces/

Working Group Meeting on New Algorithms for Inferring Molecular Structure from Distance Restraints
http://dimacs.rutgers.edu/Workshops/Molecular/

Working Group Meeting on The Burrows - Wheeler Transform: Ten Years Later
http://dimacs.rutgers.edu/Workshops/BWT/

A web page that includes open problems on embeddings of finite metric spaces, including ones generated from the Workshop on Discrete Metric Spaces and their Algorithmic Applications is
http://kam.mff.cuni.cz/~matousek/metrop.ps.gz

VIII. Contributions within Discipline

The interdisciplinary emphasis of this project has led to a variety of collaborations. Here are a few selected examples of the collaborations that have been fostered among participants.

"Lewis and Trosset are collaborating on a series of papers on this topic. One paper is in press (Computational Statistics and Data Analysis), a second has been submitted and is under review, and a third is nearing completion. The third was co-authored by Ian Grooms, then a William & Mary undergraduate, whose senior thesis project it reports. This year, we are supervising a second senior thesis project (the student is Matt Rupprecht) on the same subject, which we expect to lead to several more papers. Tarazaga wrote to me that the workshop had a profound effect on his research… I visited him last fall and he described some of the work he was doing. This will certainly lead to several papers. Last fall, Lewis, Tarazaga, and Trosset submitted a proposal to NSF to fund their collaboration. The proposal was not funded. We expect to submit a modified proposal this fall. Kearsley and Winsberg collaborated briefly on a problem that Winsberg described at the workshop. When Kearsley became too busy to continue collaborating, Winsberg teamed up with Patrick Groenen (Erasmus University Rotterdam). I've heard this work presented at a conference. In summary, …everyone considered the workshop a success and it proved to be of enormous value to three participants (Lewis, Tarazaga, Trosset). Work directly related to the mission of the workshop is ongoing."
Michael W. Trosset
College of William & Mary

"I was one of the organizers of the Working Group on The Burrows - Wheeler Transform: Ten Years Later. Contacts established during the organization of the Working Group have led to a new collaboration between Paolo Ferragina, Giovanni Manzini, Veli Mäkinen, and Gonzalo Navarro. This collaboration has produced several publications. Finally, the high quality of the results presented at the working group and the related discussions stimulated us to propose a special issue of the journal Theoretical Computer Science completely devoted to the Burrows-Wheeler Transform and its applications. The proposal has

been accepted and the special issue (edited by P. Ferragina, S. Muthukrishnan, and myself) will be published in 2006."
Giovanni Manzini
University of Piemonte Orientale

Here are a few selected examples of the contributions of the project, as seen by the participants.

"The field of metric embeddings has seen remarkable progress in the last few years. The DIMACS workshop played an important role in bringing together researchers and shaping future developments in this field. The workshop brought together theoretical computer scientists, mathematicians and computer scientists working on applications to a common forum and presented an incredible learning experience for everyone. For students, this was a wonderful opportunity to get exposure to a new rapidly developing research area. The workshop (on Discrete Metric Spaces and their Algorithmic Applications) was instrumental in identifying future directions and driving subsequent research. A list of open problems was compiled by Jiri Matousek and several of these were solved in subsequent work. These advances were directly inspired by discussions at the workshop. This field has progressed at a furious pace in the last few years and workshops such as the DIMACS workshop have played a vital role in this progress. Indeed, a follow-up workshop was held in Haifa in 2004, and with all the new results that have emerged since then, yet another workshop will be needed to understand and absorb all the latest developments!"
Moses Charikar
Princeton University

In regard to the Workshop on Discrete Metric Spaces and their Algorithmic Applications "I would explain that the field was very small at the time and is still not big, therefore any opportunity for all its practitioners to get together is very influential. In recent years this has been one of the most powerful new concepts in all of theoretical computer science and it is this small community that had done much of the work from which all later developments have emerged. Finally, the speed at which this area has advanced is simply amazing. Much of it thanks to the close ties created in such meetings and the exchange of ideas that took place there. Take my paper from the ICM and just see how many of the questions mentioned there as major open problems have been settled in a matter of three years or less. Again the meeting can be credited for helping in speeding things up this way."
Nati Linial
Hebrew University

"I attended both the meeting on Burrows-Wheeler Transform and on Discrete Metric Spaces and their Algorithmic Applications. Both were outstanding events, well organized as usual, and with a very high standard of speakers and participants. It is a tribute to the strength of DIMACS that such events can be held and supported."
Graham Cormode
Bell Labs

"My own contribution has led to a request for one, possibly two, papers for a special issue of "Theoretical Computer Science", submissions due March 2006. While preparing my workshop contribution I realized several possible future directions, which were indicated at the Workshop. Some of these are the subject of current work which I hope to submit to the TCS special issue. At present it is all work in progress…",
Peter Fenwick
Department of Computer Science,
The University of Auckland,

"I work on sparse representation and fast algorithms. My participation in the Workshop on Discrete Metric Spaces and their Algorithmic Applications leads me to understand the theoretical area better. This

was also an opportunity to communicate with more well established scholars. Now I already submitted two papers…”
Jing Zou
Princeton University

IX. Contributions -- other Disciplines

This project has been inherently interdisciplinary. This is also documented in the section on Project Findings. For example, Jing Zhou’s work on computing the Fourier representation from irregularly spaced data has wide applications in processing astrophysical and seismic data, the spectral method on adaptive grids, the tracking of Lagrangian particles, and the implementation of semi-Lagrangian methods. Ferragina, Manzini, Makinen, and Navarro’s work on succinct representations of sequences can be used for text indexing. A full-text self-index is a succinct data structure that represents a text string while supporting not only access to any character but also the efficient search for an arbitrary pattern as a substring of the indexed text.

X. Contributions -- Human Resource Development

Many of the comments in the section on Contributions within Discipline illustrate the human resource development contributions of this project. A major contribution is the impact on the research programs and careers of the participants. This project fostered new collaborations both within disciplines and among disciplines, both within academe and between academe and industry. In addition, we have the following comments from Meeyoung Cha, KAIST, a student visitor supported by this project.

“I visited DIMACS during August 2004. During the period, I attended several workshops and talks held at DIMACS and AT&T research labs. DIMACS gave me an excellent experience of meeting experts in the theory, database, and networking fields. As a result of my visit, I have been collaborating with other researchers. The collaboration continued after the visit, and our work has been acknowledged at IEEE INFOCOM as a poster. The overall work is described as a technical report at KAIST, my school. Currently, the draft is under review of a conference, and I’d like to acknowledge DIMACS in the paper if the paper gets accepted.”

XI. Contributions to Resources for Research and Education

XII. Contributions Beyond Science and Engineering