

# Adrian Dobra

Work: 206-543-8460  
Home: 425-248-7483  
Email: adobra@u.washington.edu  
Web: www.stat.washington.edu/adobra

15514 Bothell Way NE  
Apt. 202  
Lake Forest Park, WA 989155

## RESEARCH INTERESTS:

- Statistical modelling in high-dimensional spaces with mixed covariates; Stochastic computation and simulation methods (MCMC, Monte Carlo) for mining massive datasets; High performance/distributed statistical computing; Graphical models and related graph theory; Categorical data analysis in large, sparse multi-way tables; Multivariate survival analysis.
- Statistical genomics; Modelling approaches to combining genomic, clinical and sequence information; Stochastic modelling of high dimensional biological networks; Graphical association networks for exploration, visualization and summarization of gene expression data.

## APPOINTMENTS AND PROFESSIONAL EXPERIENCE:

- UNIVERSITY OF WASHINGTON Seattle, WA  
DEPARTMENT OF STATISTICS and  
DEPARTMENT OF BIOBEHAVIORAL NURSING AND HEALTH SYSTEMS and  
CENTER FOR STATISTICS AND THE SOCIAL SCIENCES  
Assistant Professor of Statistics and Nursing 9/2006 – present

- DUKE UNIVERSITY Durham, NC  
DEPARTMENT OF MOLECULAR GENETICS AND MICROBIOLOGY and  
INSTITUTE OF STATISTICS AND DECISION SCIENCES  
Research Assistant Professor 3/2003 – 9/2006
  1. Developed a novel prediction model in high-dimensional datasets with mixed covariates for distributed computing. [MPI/C]
  2. Proposed novel Metropolis-Hastings simulation approaches for inference and imputation through data augmentation in multi-way contingency tables with missing data [C].
  3. Developed a novel structural learning parallel algorithm (HdBCS) that performs covariance selection in a Bayesian framework for datasets with tens of thousands of variables. This leads to positive-definite covariance matrices even if the sample size is small. [MATLAB/MPI/C].
  4. Developed a dimensionality-reduction technique based on singular factors (metagenes) within overlapping groups of variables (genes) in large-scale (gene expression) datasets [MATLAB].
  5. Developed a user-friendly graphical tool (GraphExplore) for dynamically presenting, visualizing and interrogating large complex graphs with tens of thousands of vertices [JAVA]. Project leader.

### Others:

1. Mentored graduate students.
  2. Secured research funding by contributing to grant applications.
  3. Interviewed faculty, post-doctoral, programmer and systems administrator candidates for the Statistics, Computer Science and Molecular Genetics departments at Duke University.
- DUKE UNIVERSITY *and* Research Triangle Park, NC  
STATISTICAL AND APPLIED MATH. SCIENCES INSTITUTE (SAMSI)  
Postdoctoral Fellow 9/2002 – 2/2003  
Stochastic Computation in graphical models research group
    1. Studied the performance and scalability of MCMC methods of stochastic computation and inference in Gaussian graphical models [S+/MATLAB/C].Stochastic Computation in contingency tables research group

1. Developed sampling methods for Monte Carlo computations for exact conditional inference in discrete graphical models on multi-way contingency tables [S+/C].
- DUKE UNIVERSITY *and* Research Triangle Park, NC  
 NATIONAL INSTITUTE OF STATISTICAL SCIENCES (NISS)  
 Postdoctoral Fellow 9/2001 – 8/2003
    1. Developed and implemented a Web-based query system for releasing parts of a confidential dataset while preserving the privacy of the subjects [S+/C].
    2. Served as a consultant for United States government agencies (U.S. Census Bureau, Bureau of Transportation Statistics).
  - CARNEGIE MELLON UNIVERSITY Pittsburgh, PA  
 Research assistant 9/1998 – 8/2001
    1. Developed novel and sound statistical methods for disseminating information from large sparse multi-way contingency tables with tens of thousands of cells (e.g., United States Census data). The statistical utility as well as the risk associated with each release was estimated [S+/C].
    2. Developed a statistical framework for measuring the size of the World Wide Web utilizing a hierarchical Bayes formulation of the Rash model [S+].
  - MANAGEMENT SCIENCE ASSOCIATES Pittsburgh, PA  
 Summer internship – Statistician 5/1998 – 8/1998
    1. Developed and implemented a fast model-based record linkage system for identifying duplicate records in the company's data warehouse [S+/C]. Project leader.
    2. Devised a comprehensive set of guidelines for cleaning and enhancing the quality of the data received daily from the company's clients.
  - CARNEGIE MELLON UNIVERSITY Pittsburgh, PA  
 Teaching assistant 8/1997 – 4/1998
    1. Conducted laboratories, recitations and office hours for several undergraduate and graduate statistics courses.
    2. Assisted faculty with homework grading and mentoring students.
  - UNIVERSITY OF BUCHAREST Romania  
 DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH  
 Junior lecturer 9/1995 – 6/1997
    1. Taught operations research computing classes and recitations [S+].
    2. Taught operations research seminars and lectures.
  - OMNIS GROUP Romania  
 Programmer 7/1994 – 4/1995
    1. Developed and implemented an online RSA encryption system for transmitting confidential messages between branches of a Romanian bank [Windows/C].

**EDUCATION:**

- CARNEGIE MELLON UNIVERSITY Pittsburgh, PA  
**Ph.D. in Statistics.** Advisor: Prof. Stephen E. Fienberg 8/1997 – 8/2001  
 Thesis title: *Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables.*  
 Developed efficient algorithms based on graphical models theory for computing cell bounds as well as Markov bases that characterize spaces of contingency tables induced by an arbitrary set of fixed marginals. Generalized the *Fréchet-Hoeffding* inequality for distribution functions to decomposable graphical models which leads to closed-form expressions for cell bounds and Markov bases. When no explicit formulas exist, divide-and-conquer methods for reducing the dimensionality of the problem were devised. Distribution functions on spaces of tables were also studied. All the algorithms were specifically designed to work in parallel in a distributed computing environment and were thoroughly tested at the Pittsburgh Supercomputing Centre [UNIX/MPI/C].

- UNIVERSITY OF BUCHAREST Romania  
**M.S. in Computer Science**, Merit Fellowship, Graduated with honour (4.0 GPA). 9/1995 – 7/1996  
 Concentration: Efficient design of algorithms in distributed computing environments.  
**B.S. in Mathematics**, Merit Fellowship, Graduated with honour (3.95 GPA). 9/1990 – 7/1995  
 Concentration: Numerical algorithms, operations research.  
 Thesis title: *Maximization of a convex function over a polytope*. Advisor: Prof. V. Preda.  
 Studied various maximization techniques based on convex, polar and surface cuts. Developed a user-friendly graphical interface for the Tuy-Zwart algorithm [Windows/C].

#### PAST AND ONGOING RESEARCH SUPPORT:

- NSF (PI: Mike West)**. *Modelling of graphs, networks and trees for genomic applications: high-dimensional model search*, 2004 – 2009, \$259,740/5yrs. The proposal involves theory, methods, and computational developments for analysis and search in high-dimensional model spaces of statistical models.
- NIH/NHLBI (PI: Pascal Goldschmidt)**. *Comparative Approaches to Genomics and Complex Traits*, 2003 – 2007, \$156,944/3yrs. The proposal involves statistical research and development in cardiovascular genomics.
- NIH/NCI (PI: Joseph Nevins)**. *Integrative Cancer Biology Programs – Integration of Oncogenic Networks in Cancer Phenotype*, 2004 – 2009, \$2,308,675/5yrs. The proposal combines the resources and expertise of Duke University, the Dana Farber Cancer Institute, UT Southwestern Medical Center and the University of Southern California to focus on the development of data and computational tools to achieve an integrative and in-depth understanding of cell signalling pathways that are central to the control of cell proliferation and the oncogenic process.
- Damon Runyon-Lilly Clinical Investigator Award (PI: Jeremy Rich)**. *Maximizing Clinical Efficacy of Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors in Glioblastoma Therapy*, 2004 – 2009, \$995,000/5yrs. The goals of this project involve the correlation of signal transduction pathway activation in patient tumor samples from epidermal growth factor receptor tyrosine kinase inhibitor clinical trials with patient response and investigate the benefit of combining targeted therapies.
- NSF (PI: Alan Karr)**. *Digital Government: A Web-based Query System for Disclosure-Limited Statistical Analysis of Confidential Data*, 2001 – 2003.

#### AWARDS/FELLOWSHIPS:

- Umesh Gavasakar** Thesis Award from the Department of Statistics, Carnegie Mellon University, 2002. This award is given on an occasional basis to highly-deserving students who wrote exceptionally good Ph.D. theses.
- Carnegie Mellon University tuition fellowship (1997 – 2001).
- University of Bucharest national merit fellowship (1990 – 1995).
- Second place and silver medal at the Romanian National Mathematical Olympiad and member of the Romanian Mathematical Olympic Team (1988).
- Third place and bronze medal at the Romanian National Mathematical Olympiad (1986).
- Mention of Honour at the Romanian National Mathematical Olympiad (1985).

#### PUBLICATIONS

##### Journal papers:

- Dobra, A.**, Tebaldi, C. and West, M. (2006). *Data augmentation in multi-way contingency tables with fixed marginal totals*. Journal of Statistical Planning and Inference, 136, 355-372.
- Huber, M., Chen, Y., Dinwoodie, I., **Dobra, A.** and Nicholas, M. (2006). *Monte Carlo algorithms for Hardy-Weinberg proportions*. Biometrics, 62, 49-53.
- Chen, Y., Dinwoodie, I., **Dobra, A.** and Huber, M. (2005). *Lattice points, contingency tables, and sampling*. Contemporary Mathematics, 374, 65-78.

4. Jones, B., Carvalho, C., **Dobra, A.**, Hans, C., Carter, C. and West, M. (2005). *Experiments in stochastic computation for high dimensional graphical models*. Statistical Science, 20, 388-400.
5. DeLong, M., Yao, G., Wang, Q., **Dobra, A.**, Black, E.P., Chang, J.T., Bild, A., West, M., Nevins, J.R. and Dressman, H. (2005). *DIG – a system for gene annotation and functional discovery*. Bioinformatics, 21, 2957-2959.
6. Rich, J.N., Hans, C., Jones, B., Iversen, E.S., McClendon, R.E., Rasheed, B.K.A., **Dobra, A.**, Dressman, H.K., Bigner, D.D., Nevins, J.R. and West, M. (2005). *Gene expression profiling and analysis in graphical association studies in glioblastoma survival*. Cancer Research, 65, 4051-4058.
7. **Dobra, A.** and Sullivant, S. (2004). *A divide-and-conquer algorithm for generating Markov bases of multi-way tables*. Computational Statistics, 19, 347-366.
8. Hauser, E.R., Gregory, S., Seo, D., **Dobra, A.**, Iversen, E., Karra, R., Haynes, C., Stenger, J., Xu, H., Wang, L., Huang, L., West, M., Sketch, M., Vance, J., Kraus, W.E., Goldschmidt, P. (2004). *Convergence of genome-wide expression analysis and genome-wide linkage analysis identifies candidate genes for atherosclerosis*. Circulation 110(17:Supplement):III823.
9. **Dobra, A.**, Jones B., Hans C., Nevins J. and West, M. (2003). *Sparse graphical models for exploring gene expression data*. Journal of Multivariate Analysis, special issue on Multivariate Methods in Genomic Data Analysis, 90, 196-212.
10. **Dobra, A.** (2003). *Markov bases for decomposable graphical models*. Bernoulli, 9, No. 6, 1-16.
11. **Dobra, A.**, Karr, A. and Sanil A. (2003). *Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues*. Statistics and Computing, 13, 363-370.
12. **Dobra, A.**, Karr, A., Sanil, A. and Fienberg, S. E. (2002). *Software systems for tabular data releases*. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, special issue on Aggregation and Risk Assessment on Statistical Disclosure Control, 10, 529-544.
13. Karr, A., **Dobra, A.** and Sanil, A. (2002). *Table servers: protecting confidentiality in tabular data releases*. Communications of the ACM, special issue on Digital Government, 46, No. 1, 57-58.
14. **Dobra, A.** and Fienberg, S. E. (2001). *Bounds for cell entries in contingency tables induced by fixed marginal totals*. UNECE Statistical Journal, 18, 363-371.
15. **Dobra, A.** and Fienberg, S. E. (2000). *Bounds for cell entries in contingency tables given marginal totals and decomposable graphs*. Proceedings of the National Academy of Sciences, 97, No. 22, 11885-11892.

Book chapters:

1. **Dobra, A.**, Fienberg, S. E. and Trottni, M. (2003). *Assessing the risk of disclosure of confidential categorical data*. Bayesian Statistics 7 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.), Oxford University Press, 125-144.
2. **Dobra, A.** and Fienberg, S. E. (2003). *How large is the World Wide Web? Web Dynamics* (M. Levene and A. Poulouvassilis, eds.), Springer-Verlag, 23-45.
3. **Dobra, A.** and Fienberg, S. E. (2003). *Bounding entries in multi-way contingency tables given a set of marginal totals*. In Y. Haitovsky, H. R. Lerche and Y. Ritov, editors, Foundations of Statistical Inference, Proceedings of the Shores Conference 2000, 3-16. Springer-Verlag, Berlin.
4. **Dobra, A.**, Erosheva, E. A. and Fienberg, S. E. (2002). *Disclosure limitation methods based on bounds for large contingency tables with application to disability data*. In H. Bozdogan, ed., Statistical Data Mining and Knowledge Discovery, CRC Press.
5. **Dobra, A.**, Karr, A. and Sanil, A. (2002). *Optimal tabular releases from confidential data*. Proceedings of the National Conference for Digital Government Research, Redondo Beach, Los Angeles, CA, May 2002.

Technical reports:

1. Hans, C., **Dobra, A.** and West, M. (2005). *Shotgun stochastic search in regression with many predictors*. ISDS Discussion Paper 2005-10.

2. **Dobra, A.**, Wang, Q. and West, M. (2004). *Graphical model-based gene clustering and metagene expression analysis*. ISDS Discussion Paper 2004-24.
3. **Dobra, A.** and West, M. (2004). *Bayesian covariance selection*. ISDS Discussion Paper 2004-23.
4. Wang, Q., Yao, G., Nevins, J., West, M. and **Dobra, A.** (2004). *GraphExplore: a software tool for network visualization*. ISDS Discussion Paper 2004-22.

## PROFESSIONAL ACTIVITY

### Co-organizer:

- *Random Graphs and Stochastic Computation Workshop*. Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC, 6/2005.

### Session Chair:

- *Challenges in Stochastic Computation Workshop*. Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC, 9/2002.

### Invited talks:

1. *High-dimensional graphical models and random graphs*. Keynote speaker, Statistics for Biological Networks, EURANDOM, Eindhoven, The Netherlands, 1/2006.
2. *Bayesian covariance selection and extensions to models with mixed variables*. Joint Statistical Meetings, Minneapolis, 8/2005.
3. *Structural learning in graphical models with mixed variables*. Random Graphs and Stochastic Computation Workshop. Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC, 6/2005.
4. *Graphical association networks for exploration, visualization and summarization of gene expression data*. International Conference on Analysis of Genomic Data, Harvard Medical School, Boston, MA, 5/2004.
5. *Sparse graphical models for gene expression data*. Institute of Statistics and Decision Sciences, Duke University, Durham, NC, 10/2003.
6. *Stochastic search and optimization approaches*. Challenges in Stochastic Computation Mid-term Workdays, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC, 1/2003.
7. *Challenges in stochastic computation, and sampling posterior distributions over spaces of contingency tables*. First Cape Cod Workshop on Monte Carlo Methods, Hyannis, MA, 9/2002 (with M. West).
8. *Posterior distributions over spaces of contingency tables*. Challenges in Stochastic Computation Workshop, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC, 9/2002.
9. *Assessing the risk of disclosure of confidential categorical data*. Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, 6/2002 (with S. Fienberg and M. Trottini).
10. *Optimal tabular releases*. National Institute of Statistical Sciences, Affiliates Technology Day: Data Confidentiality, Washington D.C., 5/2002.
11. *Disclosure limitation in multi-way contingency tables*. Institute of Statistics and Decision Sciences, Duke University, Durham, NC, 3/2002.
12. *Disclosure limitation in multi-way contingency tables*. Department of Mathematics, San Francisco State University, San Francisco, CA, 2/2002.
13. *Measuring the disclosure risk for multi-way tables with fixed marginals corresponding to decomposable log-linear models*. Grostat V Conference (on applications of computational commutative algebra to statistics), New Orleans, LA, 9/2001.

### Contributed talks:

1. *Bayesian covariance selection applied to gene expression data*. International Biometric Society Eastern North American Region Meeting, Austin, Texas, 3/2005.
2. *Disclosure limitation in multi-way contingency tables*. Joint Statistical Meetings, New York City, 8/2002.
3. *Bounds for cell entries in contingency tables induced by fixed marginals*. 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, Macedonia, 3/2001 (with S. Fienberg).

Poster presentations:

1. Connelly, J.J., Wang, T., **Dobra, A.**, Jose, L., Wang, L., Huang, L., Pedersen, B., Haynes, C., Vance, J.M., Kraus, W.E., Goldschmidt-Clermont, P., Hauser, E.R., Gregory, S.G. *Identifying candidate coronary artery disease susceptibility genes through genomic convergence*. American Society of Human Genetics, Salt Lake City, 10/2005.
2. Stenger, J.E., Karra, R., Seo, D., **Dobra, A.**, Burks, S.T., Xu, H., Hauser, E.R., Iversen, E., West, M., Vance, J.M., Goldschmidt-Clermont, P.J. *GATA2 and six other transcription factors specific for cis-regulatory elements significantly over-abundant in genes differentially expressed in atherosclerotic aortas map to linkage regions in the GENECARD study*. American Society of Human Genetics, Toronto, 54:44, 10/2004.
3. Hauser, E.R., Gregory, S.G., Seo, D., **Dobra, A.**, Iversen, E., Karra, R., Haynes, C.S., Stenger, J., Xu, H., Wang, L., Huang, L., West, M., Sketch, M., Vance, J.M., Kraus, W.E., Goldschmidt, P.J. *Convergence of genome-wide expression analysis and genome-wide linkage analysis identifies candidate genes for atherosclerosis*. American Society of Human Genetics, Toronto, 54:526, 10/2004.
4. **Dobra, A.**, Tebaldi, C. and West, M. *Reconstruction of contingency tables with missing data*. Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, 6/2002.
5. **Dobra, A.** (2000). *Bounds for cell entries in contingency tables*. At the Meeting of the Pittsburgh Chapter of the American Statistical Association, University of Pittsburgh, 5/2000.
6. **Dobra, A.**, (1999). *How large is the World Wide Web?* At the Meeting of the Pittsburgh Chapter of the American Statistical Association, University of Pittsburgh, 5/1999.

**PROFESSIONAL SERVICE/AFFILIATIONS:**

1. Member of ASA.
2. Referee for JASA, JCGS, JRSS, JOS, NSF, PNAS, Statistics in Medicine.

**COMPUTATIONAL SKILLS:**

Operating Systems: Unix/Linux and Windows.  
Computer Languages: Proficient: C, MATLAB, R, S+.  
Others: PASCAL, PROLOG, LISP, SCHEME, BASIC, SQL, SAS, JAVA, EXCEL.  
Software: LATEX, WORD.  
Algorithm development: Expert-level parallel programming (MPI-based), Markov Chain Monte Carlo methods.

**OPEN-SOURCE SOFTWARE** (<http://www.stat.duke.edu/~adobra/software.htm>)

1. *MetageneCreator*—software for reducing the dimensionality in gene expression data [MATLAB].
2. *HdBCS*—software for Bayesian covariance selection in high-dimensions [C/MPI].
3. *MBCS*—a comprehensive toolbox for analyzing gene expression data using *HdBCS* [C/MATLAB].
4. *GraphExplore*—a user-friendly graphical environment for dynamically exploring large graphical structures [JAVA]. Project leader role, co-authored with Q. Wang and G. Yao.
5. *GGM*—software for stochastic computation in Gaussian graphical models [C/C++].
6. Software for data augmentation in incomplete multi-way contingency tables [C/C++].
7. Software for calculating bounds for cell entries in contingency tables with fixed marginal totals [C/C++].
8. Software for generating Markov bases using algebraic methods [C]. Co-authored with I. Dinwoodie.