

Jian Zhang

Contact information

SRI International
333 Ravenswood Ave
Menlo Park, CA 94025 Phone: +1 203-435-5280
Email: jian.zhang@sri.com
URL: <http://www.csl.sri.com/users/zhang/>

Research Interest

- Statistical machine learning
 - Hierarchical Bayes and other graphical models, model learning, inference and prediction
 - Clustering
- Scalable learning algorithms and systems for massive data sets
- Adaptive and proactive network security through machine learning
- User/customer profile modeling, collaborative filtering, recommendation and mass customization systems

Work Experience

July, 2006 – present Research Scientist, Computer Science Lab, SRI International
Sep. 2005 -- June, 2006 Postdoc Researcher, Computer Science Dept, Stanford University

Education

Ph.D. in Computer Science, Yale University 2005
Dissertation: *Massive Data Streams in Graph Theory and Computational Geometry*
Advisor: Joan Feigenbaum
M.S. in Computer Science, Yale University 2002
M.S. in Neuroscience, Brandeis University 2000
B.S. in Physiology & Biophysics, Peking (Beijing) University 1997

Projects

- **Learning hierarchical Bayesian models with nonparametric priors for user/customer profiles**
(Current project)

We construct hierarchical Bayesian models of user preferences. Nonparametric priors are used in such models. This provides great flexibility in modeling the true distribution of user preferences and leads to better predictions.

- **Adaptive and proactive cyber-defense with Gaussian process learning**
(Oct, 06 – Oct, 07, Collaborators: Phil Porras, Johannes Ullrich)
Intrusion attackers have different preferences about the attack tools they use and the networks they target. We model individual attacker preference by a preference function and model the distribution of such preference functions among the attacker population by a Gaussian process prior. The Bayesian model is learned from shared cyber-alert reports and then used to predict future attacks.
(Details in paper #1)
- **Cocustering by Similarity Refinement**
(Jan, 07 – Jun, 07)
We propose a novel framework for cocustering related objects such as documents and the words contained in these documents. In this framework, some approximate clustering structure in one type of objects (e.g. terms) is used to refine the similarity between the objects of the other type (e.g. documents). An alternating refinement process leads to better clustering of both types of objects.
(Details in paper # 2)
- **Scalable mining of correlations**
(Jan, 06 – Jun, 06, Collaborator: Joan Feigenbaum)
We designed and implemented a scalable algorithm for mining correlations among objects in extremely large data sets using min-hashing. The algorithm outperforms the next best algorithm, in terms of speed, by several orders of magnitude.
(Details in paper # 5)
- **BGP update anomaly detection by wavelet analysis**
(Sep, 04 – May, 05, Collaborators: Joan Feigenbaum, Jennifer Rexford)
We use wavelet analysis to extract features from BGP update sequences. The features are then used by a nearest neighbor learning process to identify outliers/anomalies in BGP updates.
(Details in paper # 6)
- **Efficient graph and geometry algorithms for data streams**
(2002 – 2004, Supervisor: Joan Feigenbaum)
We design data-stream algorithms for geometric and graph problems such as diameter and shortest-path distances. We also proved streaming-space lower bounds for these problems.
(Details in papers # 7, 8, 9, 10, 11)
- **Numerical simulation of neuron activities and networked neurons.**
(1999, Supervisor: Larry Abbott)

We design and implement computational models to simulate how human attention affects their vision process.
(Details in paper # 12)

Programming Skills

Highly Proficient in Languages: C/C++ (include STL), Java, Matlab
Other languages: Shell scripting, awk, SQL
Platform and Tools: Linux, gcc

Publications

(All papers are available at <http://www.csl.sri.com/users/zhang/publications>)

- 1) J. Zhang and P. Porras and J. Ullrich, “**Gaussian Process Learning for Cyber-Attack Early Warning**,” to appear in *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2008 .
- 2) J. Zhang, “**Coclustering by Similarity Refinement**,” in *Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA)*, 2007.
- 3) J. Zhang and P. Porras and J. Ullrich, “**A New Service for Increasing the Effectiveness of Network Address Blacklists**,” in *Proceedings of the 3rd Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)*, 2007.
- 4) J. Zhang and P. Porras and J. Ullrich, “**Highly Predictive Blacklisting**,” *SRI Technical Report*, 2007
- 5) J. Zhang and J. Feigenbaum, “**Finding Highly Correlated Pairs Efficiently with Powerful Pruning**,” in *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, 2006, pp. 152-161.
- 6) J. Zhang, J. Rexford and J. Feigenbaum, “**Learning-Based Anomaly Detection in BGP Updates**,” Abstract in *Proceedings of SIGCOMM Workshop on Mining Network Data*, 2005 pp. 219-220. The full paper is available at <http://www.cs.yale.edu/publications/techreports/tr1318.pdf>
- 7) J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang, “**Graph Distances in the Streaming Model: The Value of Space**,” in *Proceedings of the 16th Symposium on Discrete Algorithms (SODA)*, 2005, pp. 745 – 754.
- 8) J. Feigenbaum, S. Kannan, and J. Zhang, “**Computing Diameter in the Streaming and Sliding-Window Models**,” *Algorithmica* 41 (2005), pp. 25 - 41.

- 9) M. Elkin and J. Zhang “**Efficient Algorithms for Constructing $(1+\epsilon, \beta)$ -Spanners in the Distributed and Streaming Models,**” in *Distributed Computing* 18(5):375-385, 2006. A preliminary version appeared in *Proceedings of 23th ACM Symposium on Principles of Distributed Computing (PODC)*, 2004, pp. 160-168
- 10) J. Feigenbaum, S. Kannan, A. McGregor, S. Suri, and J. Zhang, “**On Graph Problems in a Semi-Streaming Model,**” in *Theoretical Computer Science* 348 (2005), pp. 207-216. (Special issue of selected papers from ICALP'04.) A preliminary version appeared in *Proceedings of 31th International Colloquium on Automata, Languages, and Programming (ICALP)*, LNCS Vol. 3142, 2004, pp. 531-543.
- 11) J. Feigenbaum, S. Kannan and J. Zhang “**Annotation and Computational Geometry in the Streaming Model,**” *Yale University Technical Report YALEU/DCS/TR-1249*, 2003
- 12) J. Zhang and L.F. Abbott, “**Gain Modulation in Recurrent Networks.**” In Bower, J. ed. *Computational Neuroscience*, Trends in Research 2000 (Elsevier, Amsterdam) pp. 623-628. Also in *Neurocomputing* 32-33: 623-628 (2000).

Teaching Experience

Teaching Assistant Fall 2004 CS 477
 TA for advanced undergraduate/graduate class: Neural Networks for Computing

Teaching Assistant Spring 2003 CS460
 TA for advanced undergraduate/graduate class: Theoretical Methods in Computer Science

Teaching Assistant Spring 2002 CS422
 TA for advanced undergraduate/graduate class: Operating Systems

Teaching Fellow Fall 2001 CS470
 TA for advanced undergraduate/graduate class: Artificial Intelligence

Honors

Yale University Graduate Fellowship,	2000
Brandeis University Fellowship,	1997
Pratana Scholarship, Peking University,	1996
Xinsheng Scholarship, Peking University,	1993
Zhengzhong Scholarship,	1992

References

- Prof. Joan Feigenbaum
Department of Computer Science
Yale University
P.O. Box 208285
New Haven, CT 06520-8285
Phone: (203) 432-6432
Email: Joan.Feigenbaum@yale.edu

- Prof. Avi Silberschatz
Department of Computer Science
Yale University
P.O. Box 208285
New Haven, CT 06520-8285
Phone: (203) 432-4713
Email: avi@cs.yale.edu

- Prof. Sampath Kannan
Department of Computer and Information Science
University of Pennsylvania
3330 Walnut Street
Philadelphia, PA 19104
Phone: (215) 898-9514
Email: kannan@cis.upenn.edu