



[Princeton University](#)
[Computer Science](#)
[Department](#)

Computer Science 598E

Advanced Topics in Computer Science: Cool Systems Ideas: Inspirations from Theory

Spring
2005

[Kai Li](#)

Course Summary

This goal of this graduate seminar is to identifying systems research opportunities that use practical theory results.

Students who take this course can consider it as a "crossover" seminar between systems and theory. We will be reading and discussing some selected papers in systems whose designs are heavily influenced by theory results and some in theory whose ideas are promising for designing future systems. The topics being considered include storage systems, similarity searches and data explorations. When reading theory papers, we will take a systems approach: focusing on data structures, algorithms and properties, not on how to prove the properties. We plan to explore several kinds of data including images, audio data, microarray genomic data, time series data, and network data.

Students who take this course for credits are required to present a selected paper and work on a small research project. Students can work individually or in teams. Several faculty members and outside researchers will be invited to give lectures in this seminar.

The first meeting time is 1:30pm in room 301.

Tentative Reading List and Schedule

Week 1 (2/5):

- P. Lyman, H.R. Varian, K. Swearingen, P. Charles, N. Good, L.L. Jordan, and J. Pal. [How Much Information 2003?](#)
- E. Grochowski and R. D. Halem, [Technological impact of magnetic hard disk drives on storage systems](#), IBM J. Res. & Dev 42(2), 2003.
- J. Gray; D.T. Liu; M.A. Nieto-Santisteban; A. S. Szalay; G. Heber; D. DeWitt. [Scientific Data Management in the Coming Decade](#). MSR-TR-2005-10, January 2005.

Week 2 (2/11): Fingerprinting and Systems (William Josephson)

[Lecture notes](#)

- A. Broder. [Some applications of Rabin's fingerprinting method](#). In R. Capocelli, A. De Santis and U. Vaccaro, editors, *Sequences II: Methods in Communications, Security, and Computer Science*, Springer

Verlag, 1993.

- Udi Manber. [Finding similar files in a large file system](#). In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 10–21, San Francisco, CA, USA, 1994.
- Sean Quinlan and Sean Dorward, [Venti: A New Approach to Archival Storage](#). In *Proceedings of the USENIX Conference on File And Storage Technologies (FAST)*, January 2002.
- Athicha Muthitacharoen, Benjie Chen, and David Mazières. [A Low-bandwidth Network File System](#). In *Proceedings of the ACM 18th Symposium on Operating Systems Principles*. Banff, Canada. October, 2001.
- N. T. Spring and D. Wetherall. [A protocol-independent technique for eliminating redundant network traffic](#). In *Proc. of ACM SIGCOMM*, pages 87–95, Aug. 2000.
- Frank Dabek, Emma Brunskill, M. Frans Kaashoek, David Karger, Robert Morris, Ion Stoica, and Hari Balakrishnan, [Building Peer-to-Peer Systems with Chord, a Distributed Lookup Service](#), *Proceedings of the Eighth Workshop on Hot Topics in Operating Systems (HotOS-VIII)*, May 2001.
- M. Rabin. [Fingerprinting by random polynomials](#). Report TR1581, Center for Research in Computing Technology, Harvard University, 1981.

Week 3 (2/18): Resemblance, Containment, and Min-Wise Permutations (Dr. Andrei Broder, IBM Watson)

[Lecture Notes](#)

- A. Z. Broder. [Identifying and Filtering Near-Duplicate Documents](#), COM '00: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. pages 1-10, Springer-Verlag, 2000.
- A. Z. Broder. [On the resemblance and containment of documents](#). In *Proceedings of Compression and Complexity of SEQUENCES 1997*.
- Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, Geoffrey Zweig: [Syntactic Clustering of the Web](#). *Computer Networks*. 29(8-13): 1157-1166. 1997.
- A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. [Min-Wise Independent Permutations](#). *Journal of Computer Systems and Sciences*, vol. 60(3), pp. 630-659 (2000) (special issue for STOC '98), preliminary version in *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (1998)*.
- S. Schleimer, D.S. Wilkerson and A. Aiken. [Winnowing: local algorithms for document fingerprinting](#). *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 2003. Pages 76-85.

Week 4 (2/25): Music Search and Audio Similarity (Prof. Perry Cook)

[Lecture notes](#)

- Tzanetakis, George and Perry Cook. [Musical Genre Classification of Audio Signals](#), *IEEE Transactions on Speech and Audio Processing*, 10(5), July 2002
- George Tzanetakis and Perry Cook. [MARSYAS: A Framework for Audio Analysis](#). Organized Sound, Cambridge University Press 4(3), 2000.
- G. Scavone, S. Lakatos, S., P. Cook, & C. Harbke, [Perceptual spaces for sound effects obtained with an interactive similarity rating program](#), *Intl. Symposium on Musical Acoustics*, Perugia, Italy., Sept. 2001.

- P. Cook and S. Lakatos, [Using DSP-Based Parametric Physical Synthesis Models to Study Human Sound Perception](#), Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, NY, Oct. 2003.

Week 5 (3/4): Dimension Reduction Techniques (Prof. Moses Charikar)

Lecture notes

- P. Indyk, and R. Motwani. [Approximate Nearest Neighbors: Towards Removing The Curse of Dimensionality](#). *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pages 604-613. 1998.
- E. Kushilevitz, R. Ostrovsky, and Y. Rabani. [Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces](#). *SIAM J. Comput.*, 30(2):457-474, 2000. Preliminary version appeared in STOC '98.
- M. Charikar, [Similarity Estimation Techniques from Rounding Algorithms](#). *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 380-388. 2002.
- M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni: [Locality-Sensitive Hashing Scheme Based On P-Stable Distributions](#). *Proceedings of the 20th Symposium on Computational Geometry*, pages 253-262, 2004.
- Some downloadable code for LSH is available [here](#):

Week 6 (3/11): Bloom Filters and Applications (Elliott Karpilovsky)

Lecture notes

- A. Broder and M. Mitzenmacher. [Network applications of bloom filters: A survey](#). To appear Internet Mathematics. 2005.
- Fan, L., Cao, P., Almeida, J. and Broder, A. (1998) [Summary cache: A scalable wide-area Web cache sharing protocol](#). Proceedings of ACM SIGCOMM Conference '98. Vancouver, British Columbia, Canada. August 31 - September 4, pp. ~254-265. ACM Press New York, NY, USA.
- Cohen, S., Matias, Y. [Spectral Bloom filters](#), SIGMOD 2003.
- Bernard Chazelle, Joe Kilian, Ronitt Rubinfeld and Ayellet Tal. [The Bloomier Filter: An Efficient Data Structure for Static Support Lookup Tables](#), *ACM-SIAM Symposium on Discrete Algorithms (SODA) '04, January 2004*, 30-39.
- B. Bloom. [Space/time tradeoffs in in hash coding with allowable errors](#). *CACM*, 13(7):422-426, 1970.

Week 7 (3/25): Introduction to Genomic Data (Prof. Olga Troyanskaya, Matt Hibbs)

Lecture notes (Troyanskaya), Lecture notes (Hibbs)

- Department of Energy Genome Program, [The Human Genome Project and Beyond: A Primer](#).
- S.C. Madeira and A.L. Oliveira, [Biclustering Algorithms for Biological Data Analysis: A Survey](#). *IEEE Transactions on Computational Biology and Bioinformatics*. 1(1): 24-45. 2004.

Week 8 (4/1): Time Series Data (Zhe Wang)

Lecture notes

- Tamer Kahveci and Ambuj K. Singh. [Optimizing Similarity Search for Arbitrary Length Time Series Queries](#). *IEEE Transactions on Knowledge and Data Engineering*, 16(4): 418- 433, April 2004.
- R. Agrawal, C. Faloutsos, and A. Swami, [Efficient Similarity Search in Sequence Databases](#). *Proceedings of International Conference of Foundations of Data Organization and Algorithms*. October 1993.
- K.-P. Chan and A.W.-C. Fu, [Efficient Time Series Matching by Wavelets](#). *Proceedings of Conference on Data Engineering*. March 1999.

Week 9 (4/8): Image Data Similarity Search (Frank Battagua)

Lecture notes

- Y. Rui, T. S. Huang, and S.-F. Chang. [Image retrieval: Current techniques, promising directions and open issues](#). *Journal of Visual Communication and Image Representation*, 10(4):39{62, 1999.
- A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. [Content-based image retrieval at the end of the early years](#). *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1349{1380, 2000.
- P. Indyk, and N. Thaper, [Fast Image Retrieval via Embeddings](#), *Proceedings of the 3rd International Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.
- Qin Lv, Moses Charikar, Kai Li. [Image Similarity Search with Compact Data Structures](#), in *Proc. 13th ACM Conference on Information and Knowledge Management (CIKM)*, pages 208-217, 2004.

Week 10 (4/15): Trip to DB/IR day at Columbia University

Week 11 (4/22) Indexing for Similarity Search (Christine Lv)

Lecture notes

- A. Beygelzimer, S. Kakade, and J. Langford, [Cover Trees for Nearest Neighbor](#).
- R. Krauthgamer and J. Lee. [Navigating nets: Simple algorithms for proximity search](#). Manuscript, 2003.
- A. Gionis, P. Indyk, and R. Motwani. [Similarity search in high dimensions via hashing](#). In *Proc. Int. Conf. on Very Large Data Bases*, pages 518--529, 1999.

Week 11 (4/29): Network Data Analysis using Sketches (Haakon Larsen)

- Moses Charikar, Kevin Chen, Martin Farach-Colton: [Finding frequent items in data streams](#). *Theor. Comput. Sci.* 312(1): 3-15 (2004), preliminary version in *ICALP 2002*: 693-703.
- B. Krishnamurthy, S. Sen, Y. Zhang and Y. Chen. [Sketch-based change detection: Methods, evaluation, and applications](#). In *Proc. of ACM SIGCOMM IMC* (2003).
- Cristian Estan, Stefan Savage, George Varghese, [Automatically Inferring Patterns of Resource Consumption in Network Traffic](#). In *Proceedings of ACM SIGCOMM* (2003).
- Anukool Lakhina, Mark Crovella, Christophe Diot, [Characterization of Network-Wide Anomalies in Traffic Flows](#). In *Proceedings of ACM SIGCOMM IMC* (2004).
- Robert Schweller, Ashish Gupta, Elliot Parsons, and Yan Chen, [Reversible Sketches for Efficient and Accurate Change Detection over Network Data Streams](#). *Proc. Internet Measurement Conference*, October 2004.
- Yin Zhang, Sumeet Singh, Subhabrata Sen, Nick Duffield, Carsten Lund, [Online identification of Hierarchical Heavy Hitters: Algorithms, Evaluation, and Application](#). *Proc. Internet Measurement*

Conference, October 2004.

- A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, M. J. Strauss, [QuickSAND: Quick Summary and Analysis of Network Data](#). DIMACS Technical Report 2001-43, November 2001.

Week 12 (5/9): Project presentations

Administrative Information

Regular class meets: F 1:30-4:20, Room: 301

Professor: [Kai Li](#) - 321 CS Building - 258-4637 li@cs.princeton.edu

Graduate Coordinator: [Melissa Lawson](#) - 310 CS Building - 258-5387 mml@cs.princeton.edu

Teaching Assistants: TBA
