Privacy

Swapneel Sheth
Department of Computer Science, Columbia University
swapneel@cs.columbia.edu

Candidacy Exam

Introduction and Motivation

Introduction and Motivation

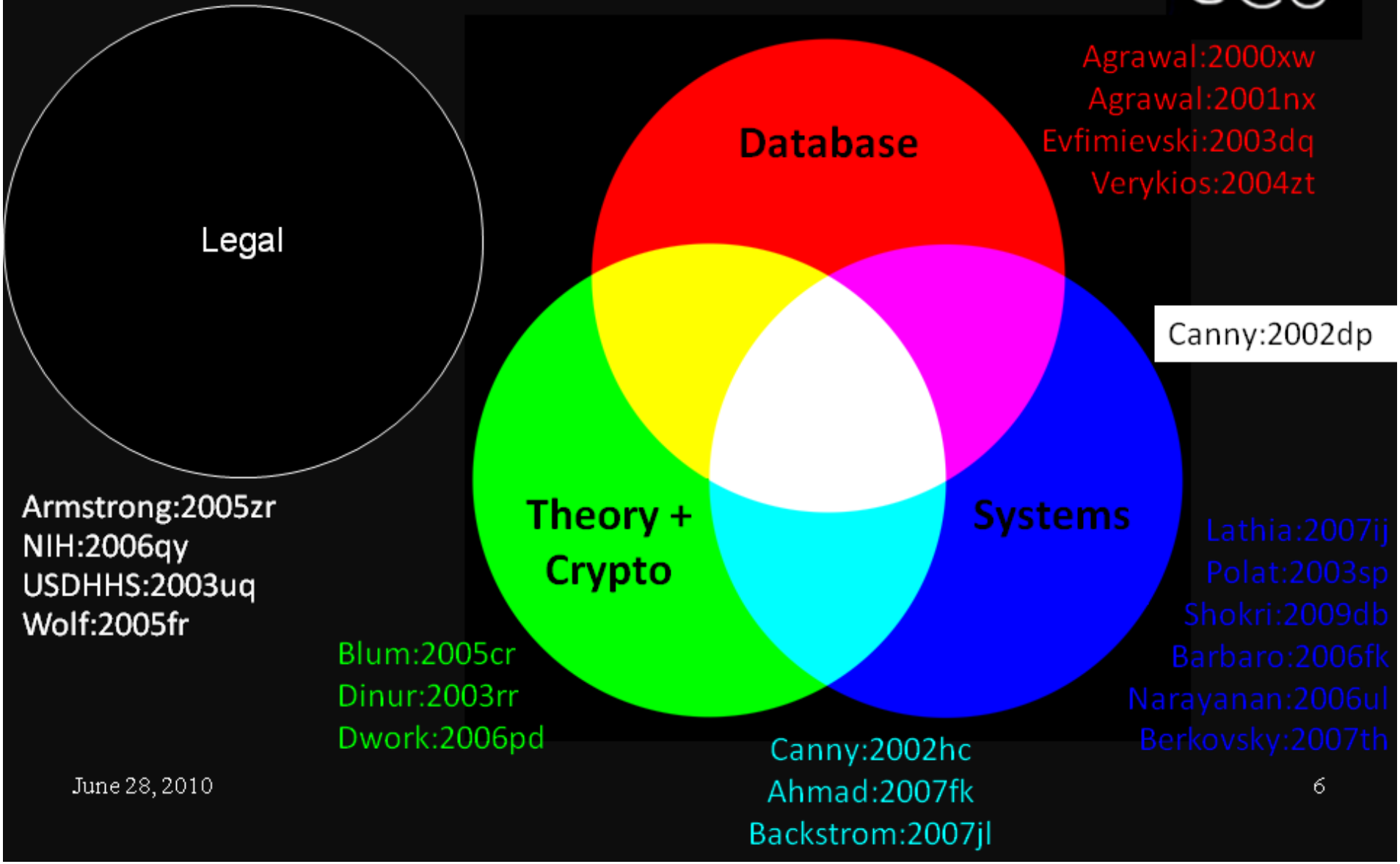
- "A Face Is Exposed for AOL Searcher No. 4417749" – [Barbaro:2006fk]
  - AOL released anonymized data for 650,000 users containing 20 million search keywords for research purposes
  - Using search history, it is possible to discern identities of the anonymized individuals
- "How To Break Anonymity of the Netflix Prize Dataset " – [Narayanan:2006ul]
  - Netflix released anonymized movie rating data for 480,000 users containing 100 millions movie ratings
  - Using public IMDB data, it is possible to identify anonymized individuals and uncover potentially sensitive information

Anonymization is not enough

Outline

- Database
- Theory + Cryptography
- Systems
- Legal
- Future Directions

Venn Diagram



Database

- "Privacy-Preserving Data Mining" – [Agrawal:2000xw]
- Introduced a quantitative measure to evaluate the amount of privacy (although later shown to be too weak [Agrawal:2001nx])
- Proposed and evaluated 3 methods for Privacy Preserving Decision-Tree Classifiers
- Preserves privacy by adding Random Perturbation to the data

[Agrawal:2000xw]

- There had been some research in the late 1970s, but it had been dormant for over 2 decades
- This paper rekindled interest in this problem in the CS community, particularly the Database community
- A lot of the later work cites this paper and tries to improve on the results

Theory + Cryptography

- "Differential Privacy" – [Dwork:2006pd]
- Shows a strong negative result – Privacy cannot be achieved if privacy is defined as "*access to a statistical database should not enable one to learn anything about an individual that could not have been possible without access*"
- This is due to "Auxiliary Information"

[Dwork:2006pd]

- Proposes an alternative definition for Privacy – "*any given privacy breach will be [...] just as likely whether or not the individual participates in the database*"
- Differential Privacy can be achieved by adding Random Noise with an exponential distribution based on the Sensitivity of the query function
- Other options exist if one wants less noise to be added (more noise, less utility) – noise can be less than sampling error provided the total number of queries is sublinear in the number of database rows [Blum: 2005cr]

Systems

- Most of the work has been in Privacy Preserving Collaborative Filtering
  - Using Randomized Perturbation Techniques (a la [Agrawal:2000xw]) – [Polat:2003sp]
  - Using Homomorphic Cryptography (a la [Canny:2002hc]) – [Ahmad:2007fk]
  - Using Distributed Aggregation of Profiles [Shokri:2009db]

Systems

- Most of the work does not use a precise definition of privacy
- Most of the work does not cite any of the recent papers in the Database or Theory communities
- Some do cite the earlier papers, but these earlier papers have later been shown to have weaknesses
- Many of the proposed solutions are not practical – e.g., [Shokri:2009db] proposes exchanging sensitive information with other users to protect the user's privacy from a malicious server
  - Most servers don't give users control over their own data
  - Need to trust the server *implicitly*

Legal

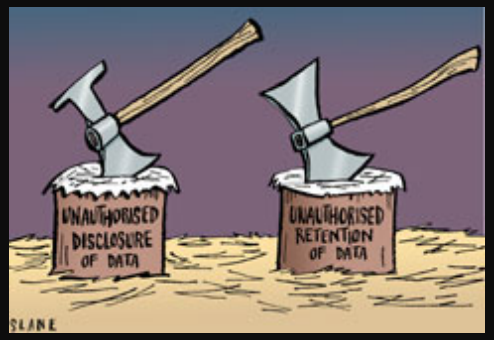
- The HIPAA Privacy Rule – [United-States-Department-of-Health-and-Human-Services:2003uq]
  - One of the *first* set of legal regulations for privacy – in this case specifically, health information
  - Defines the use and disclosure of individual's health information
  - The goal is to allow flow of health information while allowing individual's privacy
- Some privacy laws exist in other countries such as Germany

Legal

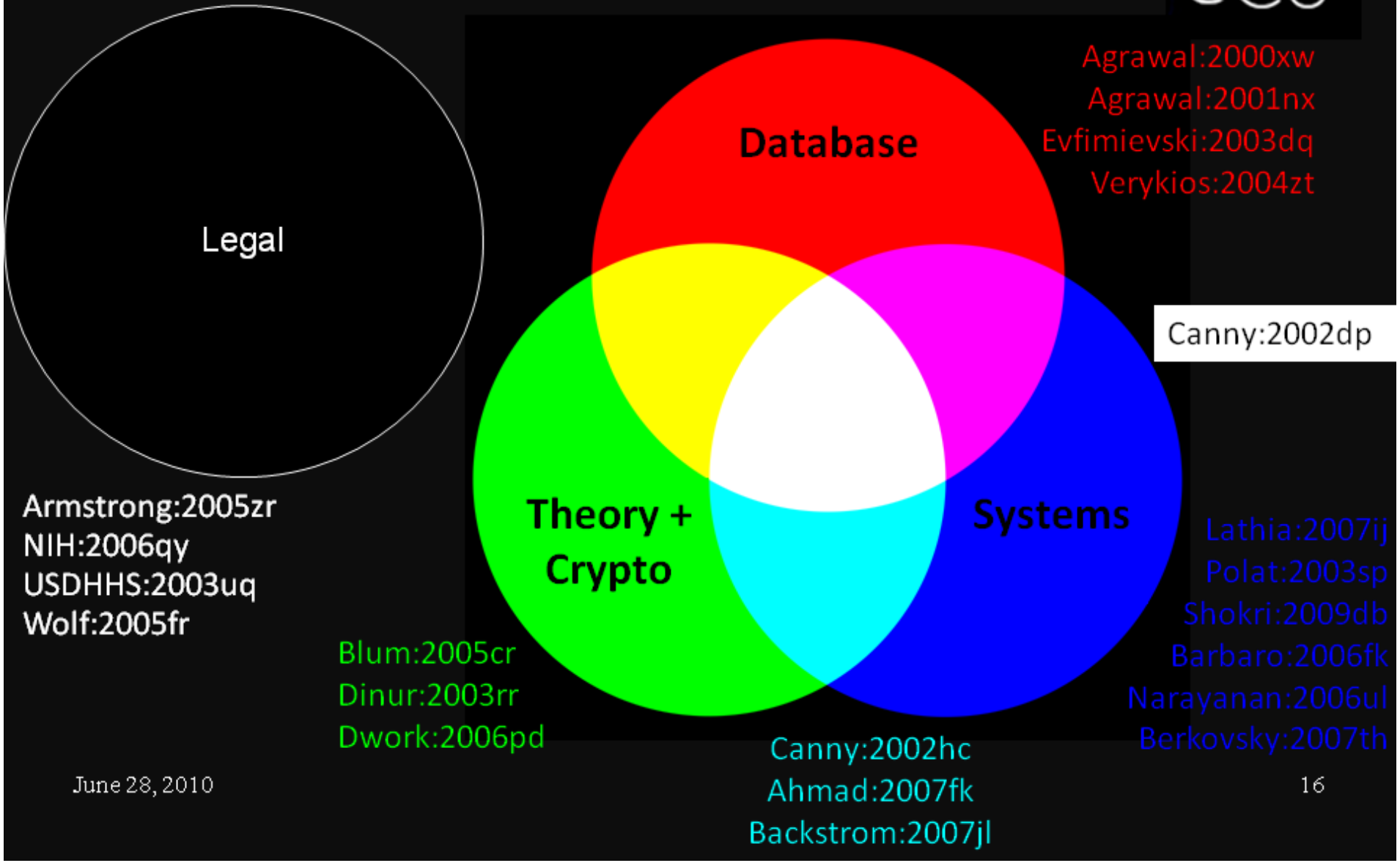
- Regulations such as HIPAA may inhibit research
- Studies [Armstrong:2005zr] [Wolf:2005fr] show
  - HIPAA increases cost and research time
  - HIPAA introduces selection bias in data collection
  - HIPAA's requirements are vague and subject to interpretation

Privacy vs OpenAccess

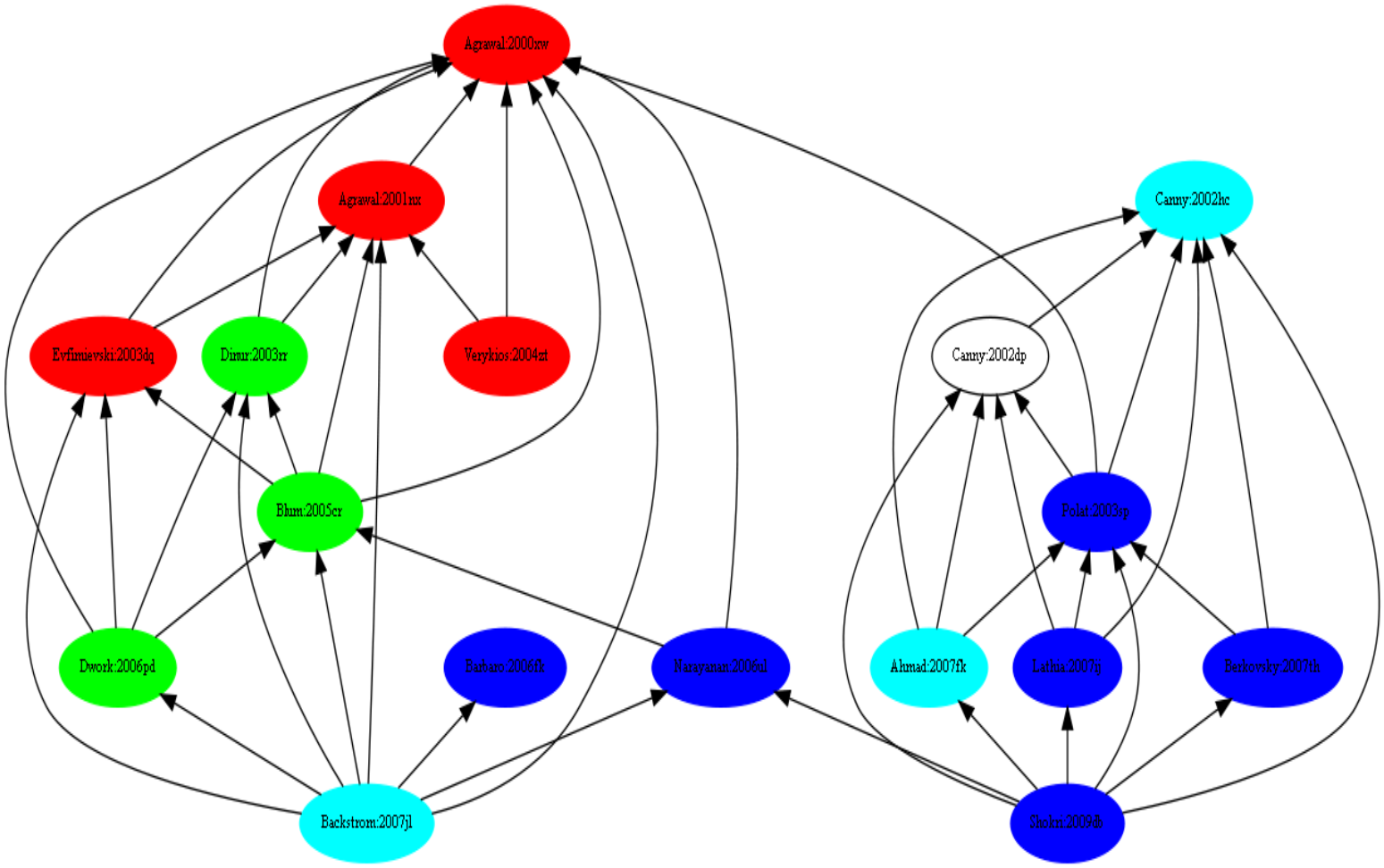
- Privacy – User has total control over his own data
- OpenAccess – Data, Publications, Software need to be publicly available for NSF/NIH funded projects
  - Beginning Oct 2010, all grant proposals need to include data management plans
  - "[…] openly sharing data will pave the way for researchers to communicate and collaborate more effectively" – Ed Seidel, NSF
  - Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans

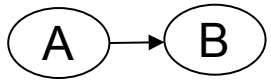


Venn Diagram

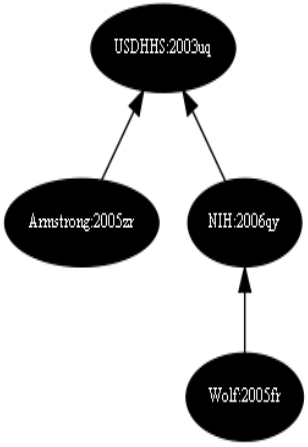


Who Cites Whom?





A cites B

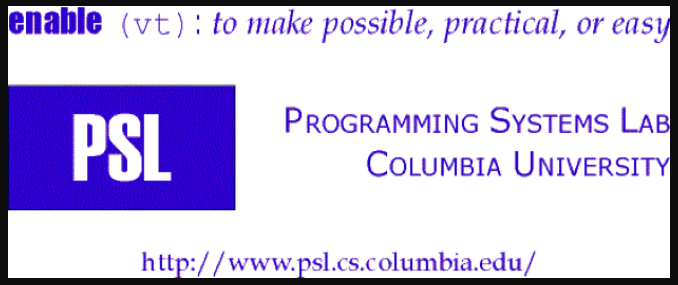


Future Directions

- Multidisciplinary Research – Databases, Theory +Crypto, Systems
- Binary vs Grayscale Data Privacy
- Client Side Caching Privacy

Future Directions (2)

- Computational Efficiency of Data Privacy – partial reusing of computation
- Energy Implications of Privacy – "Green Privacy"
- Privacy Laws and Software Localization for Privacy
- "Societal Computing" – Computing for social and legal aspects such as Privacy, Green Computing, etc.



Privacy

Swapneel Sheth
Department of Computer Science, Columbia University
swapneel@cs.columbia.edu

Bibliography

- [Agrawal:2000xw] Agrawal, R. & Srikant, R. (2000). Privacy-preserving data mining. SIGMOD Rec., 29(2), 439--450.
- [Agrawal:2001nx] Agrawal, D. & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. In PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York, NY, USA, 2001 (pp. 247--255). ACM.
- [Ahmad:2007fk] Ahmad, W. & Khokhar, A. (2007). An Architecture for Privacy Preserving Collaborative Filtering on Web Portals. In Information Assurance and Security, 2007. IAS 2007. Third International Symposium on (pp. 273-278).
- [Armstrong:2005zr] Armstrong, D., Kline-Rogers, E., Jani, S. M., Goldman, E. B., Fang, J., Mukherjee, D., Nallamothu, B. K., & Eagle, K. A. (2005). Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome. Archives of Internal Medicine, 165(10), 1125.
- [Backstrom:2007jl] Backstrom, L., Dwork, C., & Kleinberg, J. (2007). Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In WWW '07: Proceedings of the 16th international conference on World Wide Web, New York, NY, USA, 2007 (pp. 181--190). ACM.
- [Barbaro:2006fk] Barbaro, M., Zeller, T., & Hansell, S. (2006). A face is exposed for AOL searcher no. 4417749. New York Times.
- [Berkovsky:2007th] Berkovsky, S., Eytani, Y., Kuflik, T., & Ricci, F. (2007). Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems, New York, NY, USA, 2007 (pp. 9--16). ACM.
- [Blum:2005cr] Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). Practical privacy: the SuLQ framework. In PODS '05: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York, NY, USA, 2005 (pp. 128--138). ACM.
- [Canny:2002dp] Canny, J. (2002). Collaborative filtering with privacy via factor analysis. In SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2002 (pp. 238--245). ACM.
- [Canny:2002hc] Canny, J. (2002). Collaborative filtering with privacy. In Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on (pp. 45-57).

Bibliography (2)

- [Dinur:2003rr] Dinur, I. & Nissim, K. (2003). Revealing information while preserving privacy. In PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York, NY, USA, 2003 (pp. 202--210). ACM.
- [Dwork:2006pd] Dwork, C. (2006). Differential privacy. IN ICALP, 2, 1--12.
- [Evfimievski:2003dq] Evfimievski, A., Gehrke, J., & Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. In PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York, NY, USA, 2003 (pp. 211--222). ACM.
- [Lathia:2007ij] Lathia, N., Hailes, S., & Capra, L. (2007). Private distributed collaborative filtering using estimated concordance measures. In RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems, New York, NY, USA, 2007 (pp. 1--8). ACM.
- [Narayanan:2006ul] Narayanan, A. & Shmatikov, V. (2006) How to Break Anonymity of the Netflix Prize Dataset.
- [NIH:2006qy] NIH (2006). Health Services Research and the HIPAA Privacy Rule.
- [Polat:2003sp] Polat, H. & Du, W. (2003). Privacy-preserving collaborative filtering using randomized perturbation techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 625-628).
- [Shokri:2009db] Shokri, R., Pedarsani, P., Theodorakopoulos, G., & Hubaux, J.-P. (2009). Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. In RecSys '09: Proceedings of the third ACM conference on Recommender systems, New York, NY, USA, 2009 (pp. 157--164). ACM.
- [United-States-Department-of-Health-and-Human-Services:2003uq] United States Department of Health and Human Services (2003). Summary of HIPAA Privacy Rule.
- [Verykios:2004zt] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. SIGMOD Rec., 33(1), 50--57.
- [Wolf:2005fr] Wolf, M. S. & Bennett, C. L. (2005). Local perspective of the impact of the HIPAA privacy rule on research. Cancer, 106(2), 474--479.