PRIVACY PRINCIPLES AND STANDARDS

Jennifer M. Urban
UC-Berkeley School of Law
Samuelson Clinic
on behalf of EFF

Questions for the Panel

Privacy Principles and Standards:

-What are the ground rules we can all agree on? -How do we balance privacy protection with data and analytical utility?

These are the right questions. How do we answer them?

1) Draw from experience

2) Identify additional threshold questions to answer in this proceeding

3) Develop ground rules/requirements

Developing Ground Rules

Ground rules should reflect answers to the following questions

- Discussed yesterday:
  - Who needs access to the data--who is asking? Who else will ask?
  - What are their current obstacles to access?
  - What is the applicable law?
  - What are the technical security and privacy issues and constraints?
- Additional Questions:
  - What are the analytical needs? What analysis is needed?
    - What research or other questions need to be answered
  - What are data needs for any given analysis?
    - Format
    - Detail
    - Security
    - **What is *not* needed?**

Data Handling Questions from FIPPs: D. 10-06-047 June 24, 2010

Ground rules should draw on experience in this proceeding.

- D. 11-07-056 Privacy Rule reflects basic data handling questions drawn from Fair Information Practice Principles:
  - What data is the utility now collecting?
  - For what purpose is the data being collected?
  - With whom will the utility currently share the data?
  - How long will the utility currently keep the data?
  - What confidence does the utility have that the data will is accurate and reliable enough for the purposes for which the data will be used?
  - How does the utility protect the data against loss or misuse?
  - How do individuals have access to the data about themselves? and
  - What audit, oversight and enforcement mechanisms does the utility have in place to ensure that the utility is following their own rules?

Data Handling Questions from FIPPs: D. 10-06-047 June 24, 2010

- These apply equally to research and other analytical uses of energy usage data:
  - What data [does the user need to] collect[]?
    - (Level of detail, format, what is *not* needed, etc.)
  - For what purpose is the data being collected?
    - (What are the analytical needs?)

Data Minimization

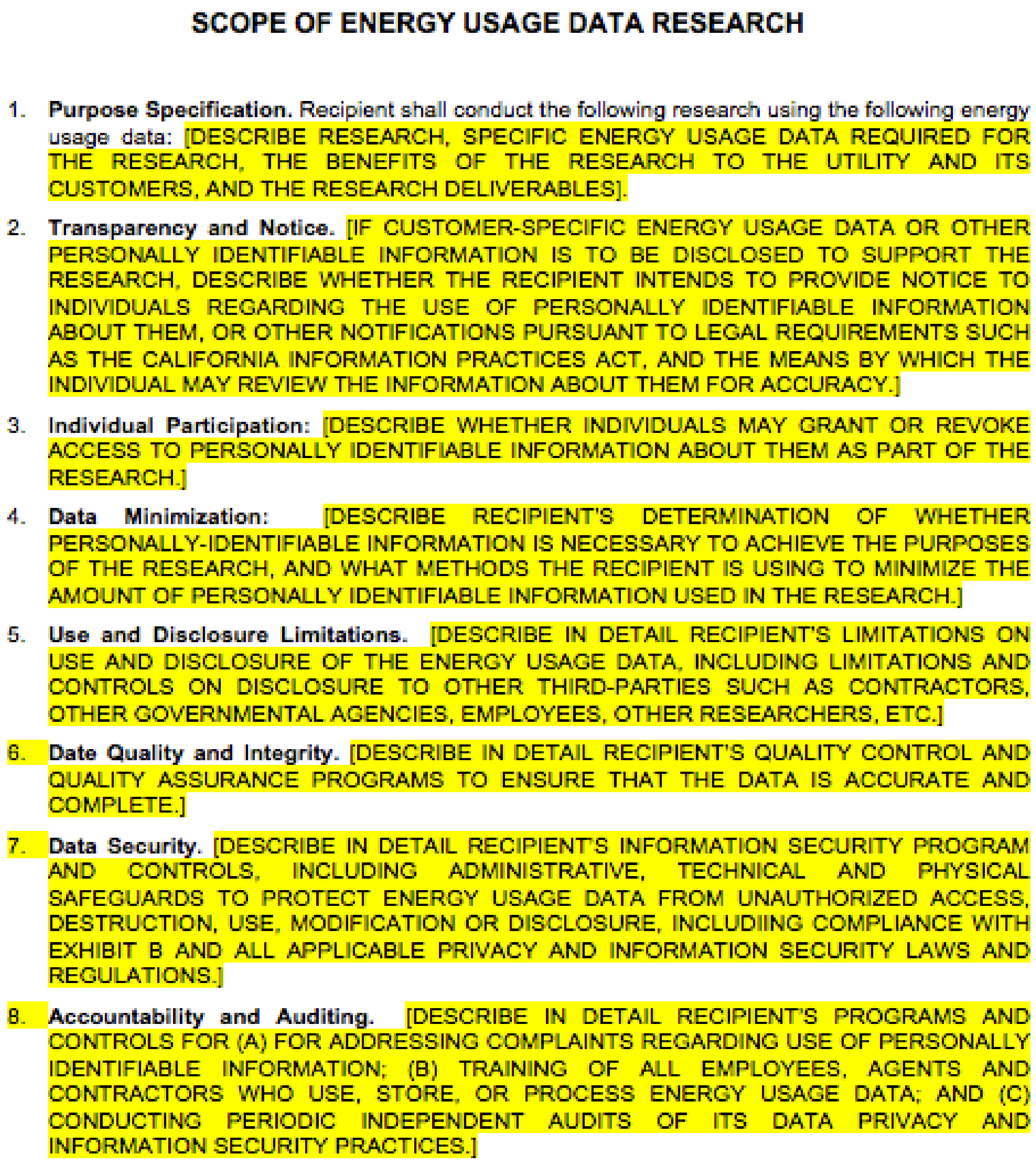
  - With whom [would the user] share the data?
  - How long [would the user need to] keep the data?
  - What confidence does the [user] have that the data [requested is what is needed] for the purposes for which the data will be used?
  - How [will the user] protect the data against loss or misuse?
  - How do individuals have access to the data about themselves? and
  - What audit, oversight and enforcement mechanisms does the [user] have in place?

2011 Privacy Rule
D. 11-07-056 July 28, 2011

Privacy Rule from D. 11-07-056 :

- Covers the threshold questions I have described
- Ground rules for research sharing will need additional detail on some issues

PG&E “Strawperson” Exhibit A



SCOPE OF ENERGY USAGE DATA RESEARCH

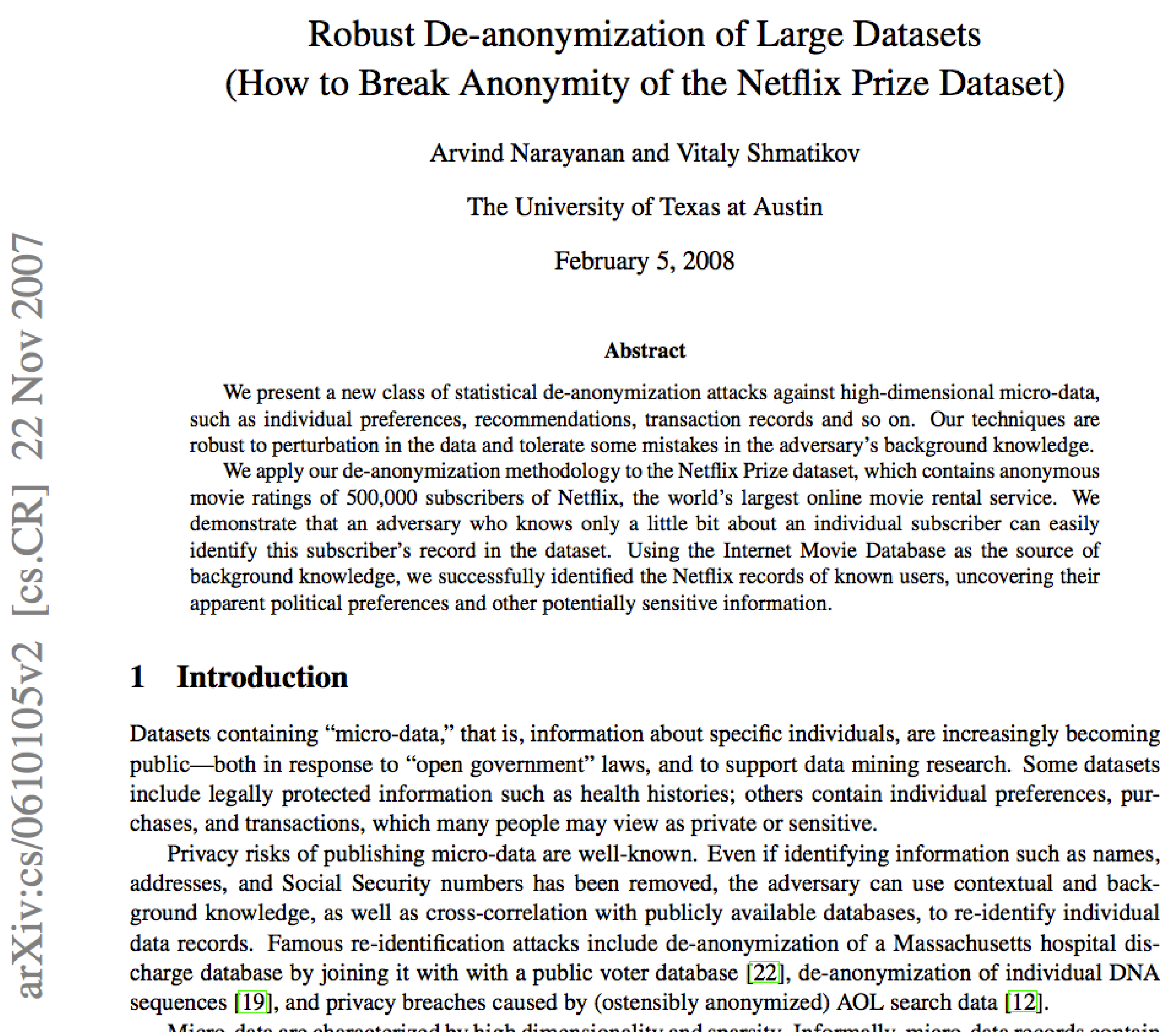
1. **Purpose Specification.** Recipient shall conduct the following research using the following energy usage data: [DESCRIBE RESEARCH, SPECIFIC ENERGY USAGE DATA REQUIRED FOR THE RESEARCH, THE BENEFITS OF THE RESEARCH TO THE UTILITY AND ITS CUSTOMERS, AND THE RESEARCH DELIVERABLES].
2. **Transparency and Notice.** [IF CUSTOMER-SPECIFIC ENERGY USAGE DATA OR OTHER PERSONALLY IDENTIFIABLE INFORMATION IS TO BE DISCLOSED TO SUPPORT THE RESEARCH, DESCRIBE WHETHER THE RECIPIENT INTENDS TO PROVIDE NOTICE TO INDIVIDUALS REGARDING THE USE OF PERSONALLY IDENTIFIABLE INFORMATION ABOUT THEM, OR OTHER NOTIFICATIONS PURSUANT TO LEGAL REQUIREMENTS SUCH AS THE CALIFORNIA INFORMATION PRACTICES ACT, AND THE MEANS BY WHICH THE INDIVIDUAL MAY REVIEW THE INFORMATION ABOUT THEM FOR ACCURACY.]
3. **Individual Participation:** [DESCRIBE WHETHER INDIVIDUALS MAY GRANT OR REVOKE ACCESS TO PERSONALLY IDENTIFIABLE INFORMATION ABOUT THEM AS PART OF THE RESEARCH.]
4. **Data Minimization:** [DESCRIBE RECIPIENT'S DETERMINATION OF WHETHER PERSONALLY-IDENTIFIABLE INFORMATION IS NECESSARY TO ACHIEVE THE PURPOSES OF THE RESEARCH, AND WHAT METHODS THE RECIPIENT IS USING TO MINIMIZE THE AMOUNT OF PERSONALLY IDENTIFIABLE INFORMATION USED IN THE RESEARCH.]
5. **Use and Disclosure Limitations.** [DESCRIBE IN DETAIL RECIPIENT'S LIMITATIONS ON USE AND DISCLOSURE OF THE ENERGY USAGE DATA, INCLUDING LIMITATIONS AND CONTROLS ON DISCLOSURE TO OTHER THIRD-PARTIES SUCH AS CONTRACTORS, OTHER GOVERNMENTAL AGENCIES, EMPLOYEES, OTHER RESEARCHERS, ETC.]
6. **Date Quality and Integrity.** [DESCRIBE IN DETAIL RECIPIENT'S QUALITY CONTROL AND QUALITY ASSURANCE PROGRAMS TO ENSURE THAT THE DATA IS ACCURATE AND COMPLETE.]
7. **Data Security.** [DESCRIBE IN DETAIL RECIPIENT'S INFORMATION SECURITY PROGRAM AND CONTROLS, INCLUDING ADMINISTRATIVE, TECHNICAL AND PHYSICAL SAFEGUARDS TO PROTECT ENERGY USAGE DATA FROM UNAUTHORIZED ACCESS, DESTRUCTION, USE, MODIFICATION OR DISCLOSURE, INCLUDIING COMPLIANCE WITH EXHIBIT B AND ALL APPLICABLE PRIVACY AND INFORMATION SECURITY LAWS AND REGULATIONS.]
8. **Accountability and Auditing.** [DESCRIBE IN DETAIL RECIPIENT'S PROGRAMS AND CONTROLS FOR (A) FOR ADDRESSING COMPLAINTS REGARDING USE OF PERSONALLY IDENTIFIABLE INFORMATION; (B) TRAINING OF ALL EMPLOYEES, AGENTS AND CONTRACTORS WHO USE, STORE, OR PROCESS ENERGY USAGE DATA; AND (C) CONDUCTING PERIODIC INDEPENDENT AUDITS OF ITS DATA PRIVACY AND INFORMATION SECURITY PRACTICES.]

“Covered Information” and “Aggregate” Information

- **Key pair of provisions in D. 11-07-056 Privacy Rule**

Energy usage data is “**covered information**” if “an individual, family, household or residence, or non-residential customer can[] ***reasonably be identified or re-identified***” (emphasis added)

Covered entities must “permit the use of **aggregated usage data** (removed of all PII) …. ***provided that the release of that data does not disclose or reveal specific customer information*** because of the size of the group, rate classification, or nature of the information”



Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

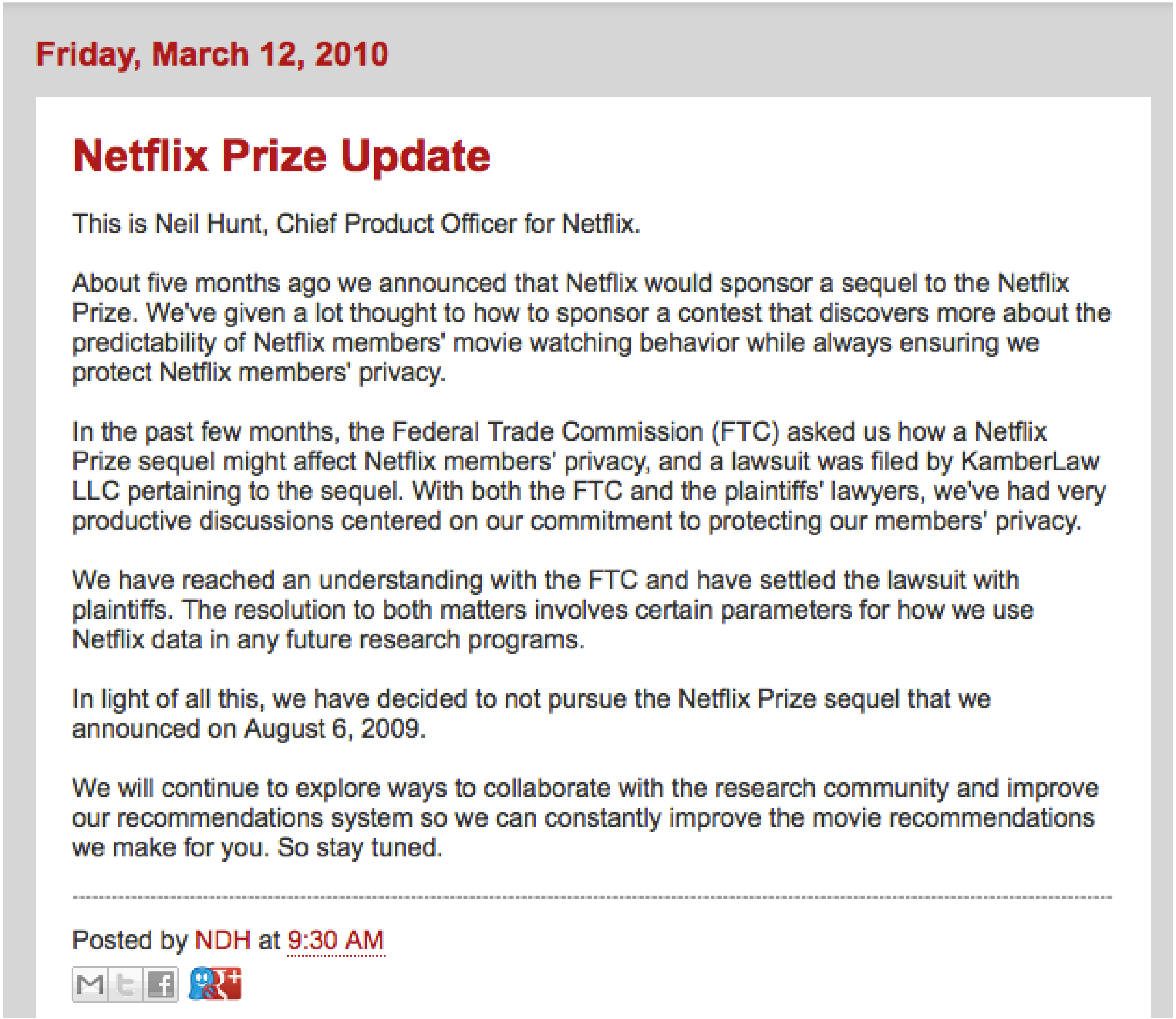
We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

1 Introduction

Datasets containing "micro-data," that is, information about specific individuals, are increasingly becoming public—both in response to "open government" laws, and to support data mining research. Some datasets include legally protected information such as health histories; others contain individual preferences, purchases, and transactions, which many people may view as private or sensitive.

Privacy risks of publishing micro-data are well-known. Even if identifying information such as names, addresses, and Social Security numbers has been removed, the adversary can use contextual and background knowledge, as well as cross-correlation with publicly available databases, to re-identify individual data records. Famous re-identification attacks include de-anonymization of a Massachusetts hospital discharge database by joining it with with a public voter database [22], de-anonymization of individual DNA sequences [19], and privacy breaches caused by (ostensibly anonymized) AOL search data [12].

Micro-data are characterized by high dimensionality and sparsity. Informally, micro-data records contain



Friday, March 12, 2010

Netflix Prize Update

This is Neil Hunt, Chief Product Officer for Netflix.

About five months ago we announced that Netflix would sponsor a sequel to the Netflix Prize. We've given a lot thought to how to sponsor a contest that discovers more about the predictability of Netflix members' movie watching behavior while always ensuring we protect Netflix members' privacy.

In the past few months, the Federal Trade Commission (FTC) asked us how a Netflix Prize sequel might affect Netflix members' privacy, and a lawsuit was filed by KamberLaw LLC pertaining to the sequel. With both the FTC and the plaintiffs' lawyers, we've had very productive discussions centered on our commitment to protecting our members' privacy.

We have reached an understanding with the FTC and have settled the lawsuit with plaintiffs. The resolution to both matters involves certain parameters for how we use Netflix data in any future research programs.

In light of all this, we have decided to not pursue the Netflix Prize sequel that we announced on August 6, 2009.

We will continue to explore ways to collaborate with the research community and improve our recommendations system so we can constantly improve the movie recommendations we make for you. So stay tuned.

Posted by NDH at 9:30 AM

“Covered Information” and “Aggregate” Information

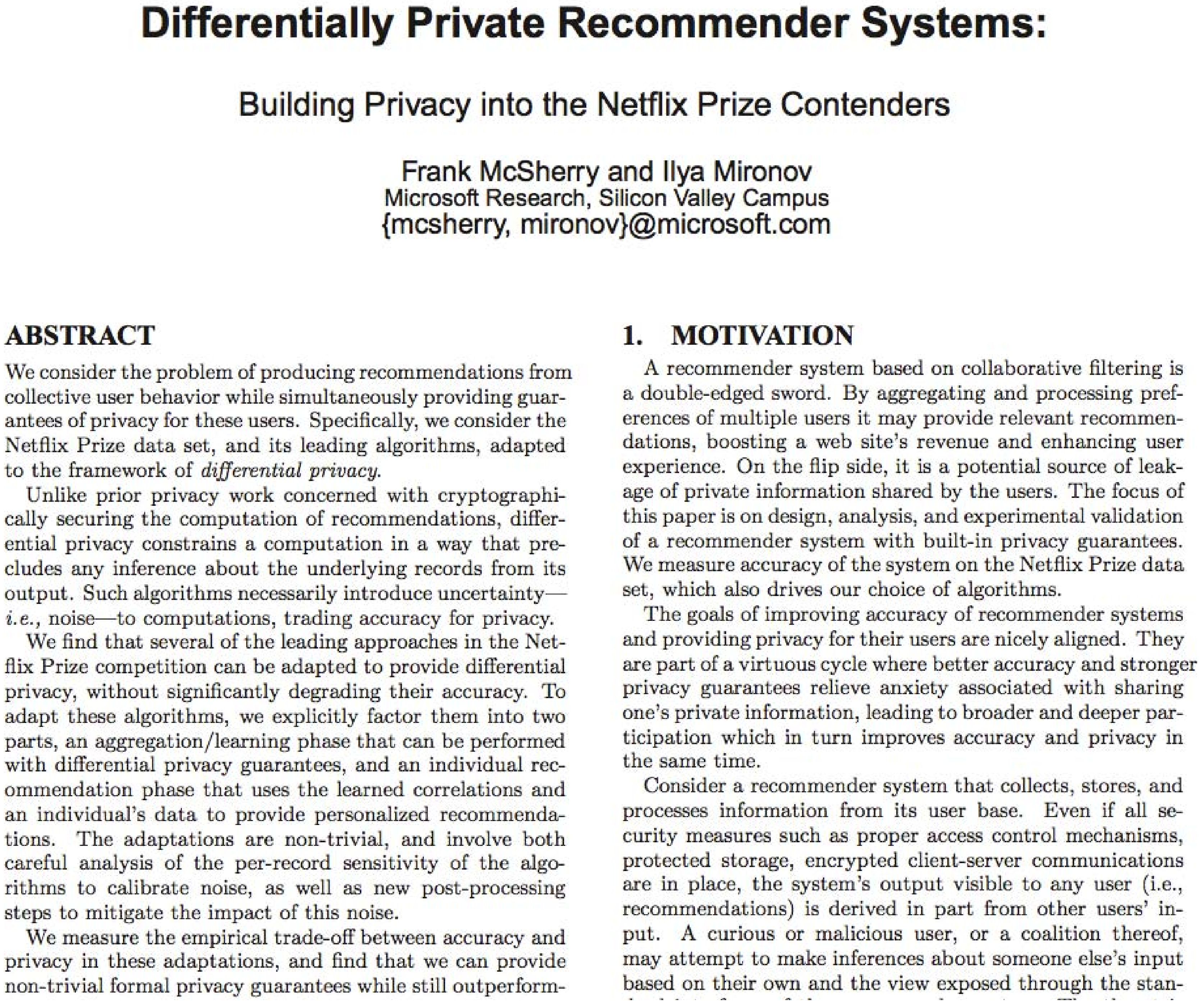
Ground Rules should specifically address:

- Re-identification of individual household or customer and revelation of “specific customer information”
  - To apply “covered information” definition
  - To apply “aggregate” information provision
- Ground Rules should require, at a minimum,
  - Policy safeguards
  - Technical best practices
  - That
    - Minimize revelation of data
    - Are developed in consultation with experts
    - Can evolve over time

Addressing Aggregation, Re-identification, and Research Needs

Ground Rules should:

- Work as a more detailed application of the Privacy Rule
- Require minimization of data revelation
- Draw from experience
  - Existing research examples
    - See comments by Haas researchers
    - Others
- Reflect expert input
  - Data security
  - Re-identification
  - Other “Big Data” privacy issues
    - Keeping data over time (multiple studies)
    - Updating practices and datasets as needed



Differentially Private Recommender Systems:

Building Privacy into the Netflix Prize Contenders

Frank McSherry and Ilya Mironov
Microsoft Research, Silicon Valley Campus
{mcsherry, mironov}@microsoft.com

ABSTRACT

We consider the problem of producing recommendations from collective user behavior while simultaneously providing guarantees of privacy for these users. Specifically, we consider the Netflix Prize data set, and its leading algorithms, adapted to the framework of *differential privacy.*

Unlike prior privacy work concerned with cryptographically securing the computation of recommendations, differential privacy constrains a computation in a way that precludes any inference about the underlying records from its output. Such algorithms necessarily introduce uncertainty—*i.e.,* noise—to computations, trading accuracy for privacy.

We find that several of the leading approaches in the Netflix Prize competition can be adapted to provide differential privacy, without significantly degrading their accuracy. To adapt these algorithms, we explicitly factor them into two parts, an aggregation/learning phase that can be performed with differential privacy guarantees, and an individual recommendation phase that uses the learned correlations and an individual's data to provide personalized recommendations. The adaptations are non-trivial, and involve both careful analysis of the per-record sensitivity of the algorithms to calibrate noise, as well as new post-processing steps to mitigate the impact of this noise.

We measure the empirical trade-off between accuracy and privacy in these adaptations, and find that we can provide non-trivial formal privacy guarantees while still outperform-

1. MOTIVATION

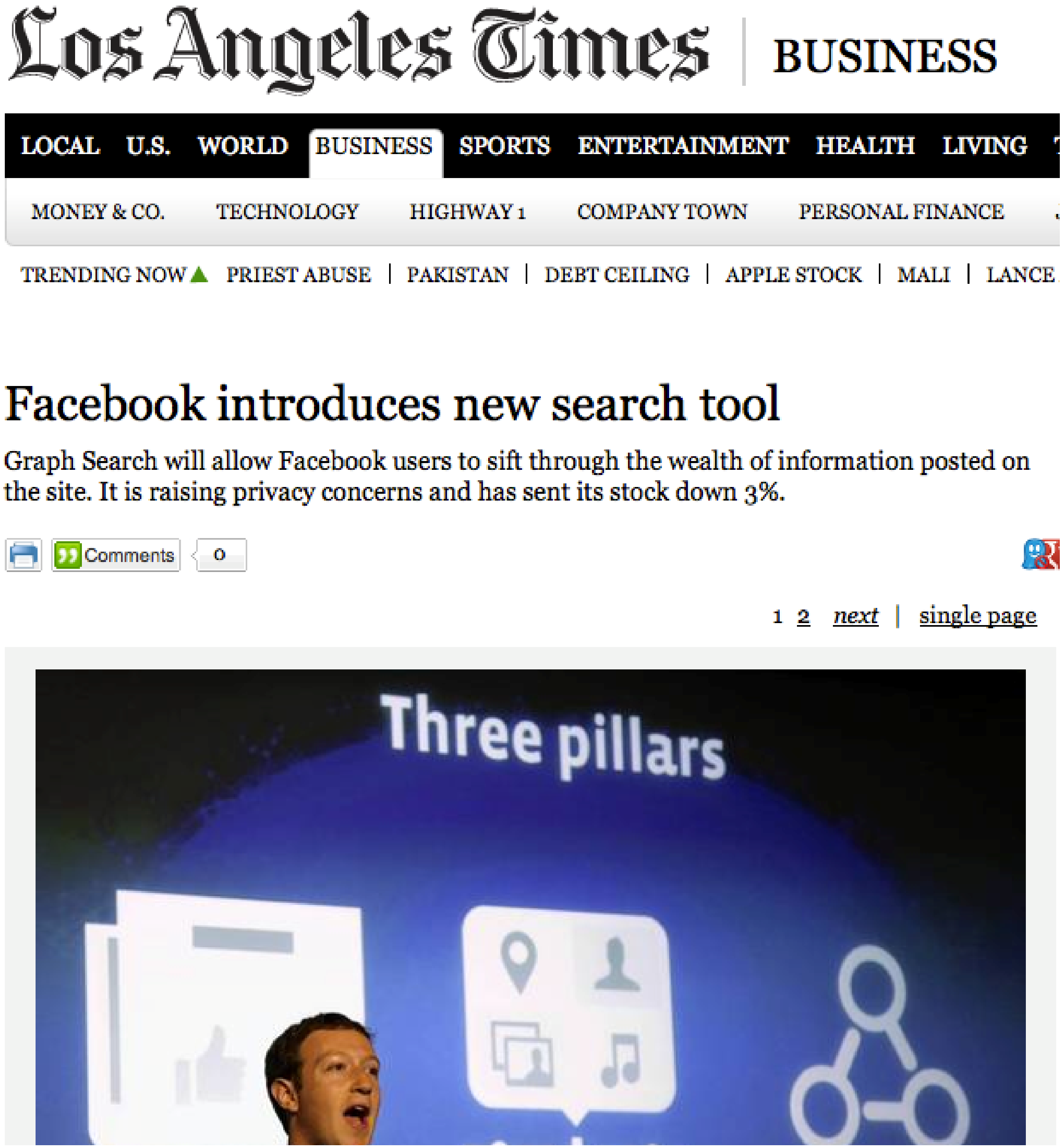
A recommender system based on collaborative filtering is a double-edged sword. By aggregating and processing preferences of multiple users it may provide relevant recommendations, boosting a web site's revenue and enhancing user experience. On the flip side, it is a potential source of leakage of private information shared by the users. The focus of this paper is on design, analysis, and experimental validation of a recommender system with built-in privacy guarantees. We measure accuracy of the system on the Netflix Prize data set, which also drives our choice of algorithms.

The goals of improving accuracy of recommender systems and providing privacy for their users are nicely aligned. They are part of a virtuous cycle where better accuracy and stronger privacy guarantees relieve anxiety associated with sharing one's private information, leading to broader and deeper participation which in turn improves accuracy and privacy in the same time.

Consider a recommender system that collects, stores, and processes information from its user base. Even if all security measures such as proper access control mechanisms, protected storage, encrypted client-server communications are in place, the system's output visible to any user (i.e., recommendations) is derived in part from other users' input. A curious or malicious user, or a coalition thereof, may attempt to make inferences about someone else's input based on their own and the view exposed through the stan-

Practical Considerations

- Practical security and privacy
  - Breaches happen
  - “Mission creep” also happens
  - Re-identification happens
- Proactive planning/design (PbD) vs. retrofitting
  - Costs: not just money, though that, too
- Backlash and mid- to longterm effects of problems
  - Maintaining versus losing the public trust (U.S. Census)
    - Applies equally to any research subjects (Boston College oral history)
    - Also applies to private companies
  - Good opportunity to avoid this



Facebook introduces new search tool

Graph Search will allow Facebook users to sift through the wealth of information posted on the site. It is raising privacy concerns and has sent its stock down 3%.

Next Steps

- Understand and apply legal backdrop
  - Information Practices Act; other legal requirements
- Gather information
  - Use FIPPs-based and Additional Questions to gather information during this proceeding
  - Draw on experience
    - From proceeding/Privacy Rule
    - From researchers and research institutions
      - E.g., institutional review board processes
    - From security experts
    - From utilities and other data custodians and handlers
- Develop requirements for data-sharing
  - That follow Privacy Rule, applicable laws, and technical best practices for research and security