

# Private Data Mining and Citizen's Rights

**Andrew Lindell**

Chief Cryptographer  
Aladdin Knowledge Systems

Assistant Professor  
Bar-Ilan University

## The Information Explosion

- **Huge databases exist in society today**
  - Government-collected data on citizens and non-citizens
  - Medical data
  - Consumer purchase data
  - Census data
  - Communication and media-related data



- **Data utilization is critical**
  - Governments use information on citizens and non-citizens for homeland security
    - Find potential terrorists
    - Law enforcement
  - Medical data is utilized for medical research
  - Businesses use consumer data for market research
  - And much more...

- **The amount of data now being collected cannot be analyzed manually**
- **Data mining algorithms are used to automatically extract high-level information and patterns**
  - Data mining algorithms have been proven successful in a wide range of applications, including homeland security

- **Investigation at Stillwater State Correctional Facility, Minnesota**
  - Data mining software was applied to phone records from the prison
  - A pattern linking calls between prisoners and a recent parolee was discovered
  - The calling data was then mined again together with records of prisoners' financial accounts
  - **The result:** a large drug smuggling ring was uncovered

- **Privacy concerns**
  - Data must be protected from exposure
  - Data must be used in appropriate ways (according to well-defined privacy policies and/or the law)
- **This suffices for the simple case where the data is used by the same organization collecting it**
  - Notice that this is **not** the case in the prison example



- **Data sources are pooled in order to achieve higher utility**
  - Pooling medical data can improve the quality of medical research
  - Pooling information from different government agencies can provide a wider picture
    - What is the health status of citizens that are supported by social welfare?
    - Are there citizens that receive simultaneous support from different agencies?

- **Many different security agencies coexist**
- **These agencies are hesitant to share information**
  - This is often justified
    - If all agencies share all information, a single mole can compromise all agencies
    - “If you have one gigantic database, you have one gigantic target for the terrorists and the bad guys”, Peter Swire
- **But more patterns could be found if data and not just conclusions are shared**

- **A head-on collision**
  - We need to be able to utilize data from different sources
  - We cannot ignore privacy concerns of law-abiding citizens
- **It seems that we have to make a choice**  
**Data mining OR privacy**



Congressional Record: July 14, 2003 (Senate)  
Page S9339-S9354

DEPARTMENT OF DEFENSE APPROPRIATIONS ACT, 2004

SA 1217. Mr. STEVENS proposed an amendment to the bill H.R. 2658, making appropriations for the Department of Defense for the fiscal year ending September 30, 2004, and for other purposes; as follows:

[...]

Sec. 8120.

- (a) Limitation on Use of Funds for Research and Development on Terrorism Information Awareness Program.-- Notwithstanding any other provision of law, no funds appropriated or otherwise made available to the Department of Defense, whether to an element of the Defense Advanced Research Projects Agency or any other element, or to any other department, agency, or element of the Federal Government, may be obligated or expended on research and development on the Terrorism Information Awareness program.
- (b) Limitation on Deployment of Terrorism Information Awareness Program.--(1) Notwithstanding any other provision of law, if and when research and development on the Terrorism Information Awareness program, or any component of such program, permits the deployment or implementation of such program or component, no department, agency, or element of the Federal Government may deploy or implement such program or component, or transfer such program or component to another department, agency, or element of the Federal Government, until the Secretary of Defense-- (A) notifies Congress of that development, including a specific and detailed description of-- (i) each element of such program or component intended to be deployed or implemented; and

[...]

- (1) **the Terrorism Information Awareness program should not be used to develop technologies for use in conducting intelligence activities or law enforcement activities against United States persons without appropriate consultation with Congress or without clear adherence to principles to protect civil liberties and privacy;**



## The Big Brother Database

OTTAWA, ONTARIO - The Minister of Human Resources Development Canada, the Honourable Jane Stewart, announced today that following discussions with the Privacy Commissioner, HRDC's information databank for labour market and social programs, the Longitudinal Labour Force File (LLFF), is being dismantled.

With the dismantling of the LLFF, HRDC has eliminated the computer program used to link its information with information from the Canada Customs and Revenue Agency and data on social assistance from provincial/territorial governments.

LLFF information from the Canada Customs and Revenue Agency has been returned to that Agency. HRDC will review the information-sharing arrangements it has with provincial and territorial governments for research purposes. The Department's policy analysis and research data relating to its own programs will be kept as separate, secure and unlinked files; all personal information identifying individuals will remain encrypted.

"The Privacy Commissioner fully supports this decision, and the other measures we are taking to protect privacy," said Minister Stewart. "In a letter to my department Mr. Phillips has said that he accepts and supports these measures, and that they satisfy all the recommendations and observations outlined in his 1999-2000 Annual Report."

**"The Privacy Commissioner acknowledges that there has never been a known breach of security with regard to this databank, and HRDC has been acting within the existing Privacy Act. However, given public concerns about privacy issues in this era of advanced and constantly changing technology, I have chosen an approach that addresses future threats to privacy."**



Andrew Lindell  
Private Data Mining and Citizens' Rights  
April 29, 2008



## Privacy-Preserving Data Mining

- **Have your cake and eat it too**
- **Enable two or more organization to jointly mine their data without revealing anything but the results**
- **Caveat**
  - This does not solve the problem of inappropriate use
  - Essentially, back to case of a single user

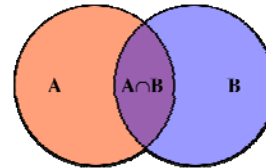


Andrew Lindell  
Private Data Mining and Citizens' Rights  
April 29, 2008



- **Set Intersection**

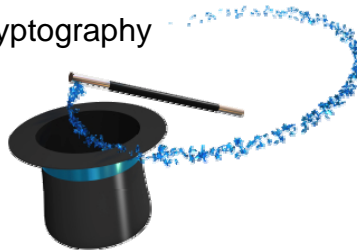
- **Input:** two or more parties with private databases (keyed by some attribute, say SSN)
- **Output:** the keys that appear in both databases (e.g., social security numbers appearing in both), and *nothing more*



- **Find people on the suspected terrorist list of different agencies**
  - While keeping each list **private**
- **Find citizens who receive social welfare but report earnings above the maximum allowed**
  - Without revealing who is on social welfare to the IRS and vice versa
- **Determine airline passengers on the government's "no fly" list**
- **Find which patients on social welfare have cancer**

- **Aren't we overdoing it on privacy?**
  - Terror is about saving lives and this is much more important!
- **Answers**
  - **Philosophical:** there is something ironic about protecting freedom by partially taking it away
  - **Practical 1:** by using privacy-preserving solutions we prevent the backlash that prevents the use of information
    - **Privacy protection enables information flow!**
  - **Practical 2:** Europe has different privacy regulations and the US needs to interact with them

- **How is it possible to compute the intersection without at least one side seeing all of the data?**
  - The magic of cryptography



- **Pseudorandom functions**
  - A **random function** is a function that assigns a random output to every input (independently of all others)
  - A **pseudorandom function** is a cryptographic function that looks like a random one
- **For simplicity, can think of this as an encryption function (but this is technically incorrect)**
  - It is correct for block ciphers like 3DES and AES
  - To be concrete we will refer to 3DES from here on

- **Input:** a set  $X = \{x_1, \dots, x_n\}$  held by Alice, and a set  $Y = \{y_1, \dots, y_n\}$  held by Bob
- **Protocol – step 1:**
  - Alice chooses a secret key  $k$  and computes the set  $X_{Enc} = \{3DES_k(x_1), \dots, 3DES_k(x_n)\}$
  - Alice sends  $X_{Enc}$  to Bob
  - Note: Bob learns nothing from  $X_{Enc}$  because he doesn't know the secret key  $k$

- **Protocol – step 2:**
  - Bob learns the values  $3DES_k(y_1), \dots, 3DES_k(y_n)$  without learning the key  $k$  (and without Alice learning  $y_1, \dots, y_n$ )
  - Bob knows that  $y_i$  is in the intersection if  $3DES_k(y_i)$  is in  $X_{enc}$  (recall that  $X_{enc} = \{3DES_k(x_1), \dots, 3DES_k(x_n)\}$ )
- **Question:**
  - Why is it important that Bob doesn't learn  $k$ ?
    - Given  $k$ , Bob can decrypt each  $3DES_k(x_i)$  and learn  $x_i$

- **Problem:**
  - How can Bob learn  $3DES_k(y_1), \dots, 3DES_k(y_n)$  without him learning  $k$ , or Alice learning  $y_1, \dots, y_n$
  - It looks like a paradox: either Alice must send  $k$  to Bob or Bob must send  $y_1, \dots, y_n$  to Alice



- **Cryptographic protocols**
  - There exist protocols for carrying out such tasks
  - The parties learn the output only, and nothing about each other's inputs
  - Such protocols are typically expensive (but progress has been made and some are "reasonable")

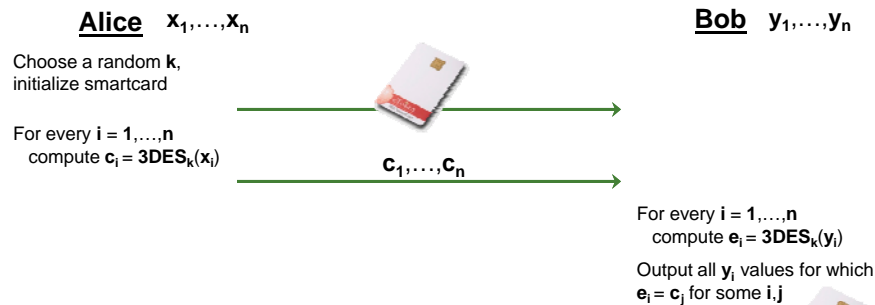
- **We will present an alternative that uses smartcard technology**



- **A smartcard is a secured piece of hardware with well-defined functionality**
- **Smartcards store cryptographic keys and can carry out operations on-board**
  - The keys never leave the smartcard
- **Smartcards have strong physical protection**
  - Self-destruct if exposed to light, or if triggered
  - Obfuscated logic
  - Miniaturization to make reverse engineering hard
  - And much much more...

- **The computation is carried out by the parties communicating over a network**
- **In addition, one of the parties prepares a standard smartcard (in some way) and physically sends it to the other**
  - Standard smartcard is important for ease of deployment
  - Standard smartcard is important for trust!
- **This model is suitable for applications of homeland security and interaction between government agencies**

- **In step 1, Alice carries out the following steps**
  - Alice initializes a smartcard with a secret key  $k$  for 3DES
  - Alice computes  $X_{Enc} = \{3DES_k(x_1), \dots, 3DES_k(x_n)\}$  on her PC
  - Alice sends the **smartcard** and  $X_{Enc}$  to Bob
- **In step 2, Bob does the following**
  - Bob computes  $Y_{Enc} = \{3DES_k(y_1), \dots, 3DES_k(y_n)\}$  using the **smartcard** (Bob does not know  $k$ )
  - Bob outputs every  $y_i$  for which  $3DES_k(y_i)$  is in  $X_{Enc}$



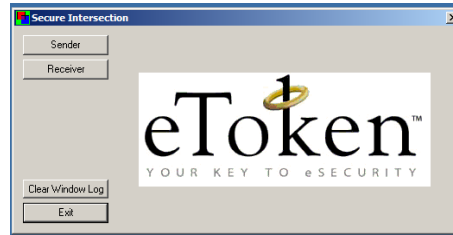
\* There is a small fix required for this protocol, as we will see soon.

- **Alice learns nothing about Bob's input**
  - Alice receives nothing from Bob so this is clear
- **Bob learns the intersection but nothing more**
  - Bob receives the list  $c_1, \dots, c_n$  which is  $3DES_k(x_1), \dots, 3DES_k(x_n)$
  - If it queries the smartcard on  $y = x_i$  then it knows that Alice has  $x_i$ . However, this is allowed because it means that  $x_i$  is in the intersection!
  - If Bob does not query the smartcard on  $x_j$ , then Bob learns nothing about  $x_j$  from  $3DES_k(x_j)$  because  $3DES$  is a pseudorandom function.

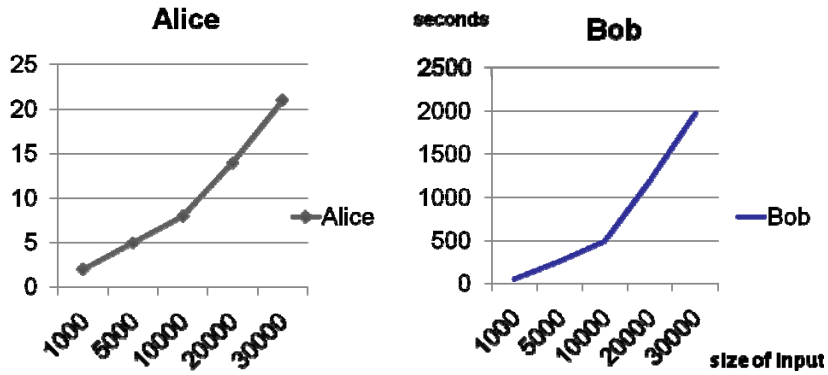
- **What prevents Bob from querying the smartcard on a huge number of values (to run an exhaustive search)?**
  - Smartcard objects can be initialized with a “usage counter” limiting the number of times an object can be used
  - When Alice initializes the smartcard with a  $3DES$  key, she sets the usage counter to  $n$  (or whatever the size of Bob's input set)

- **Highly efficient**
  - Alice carries out all 3DES operations on a PC
  - Bob computes 1 smartcard 3DES operation per input value
    - At 50ms per operation, we get 72000 in one hour
    - Our implementation works at approximately this rate
- **Provable security**
  - The protocol can be proven secure under stringent definitions, demonstrating that nothing beyond the set intersection itself can be learned
- **Simple to Implement**

- **We implemented the protocol using Aladdin's eToken PRO**
  - No attempt has been made to optimize the code
  - Nevertheless, it is very efficient
  - For **10,000** records (using an IBM T41p laptop)
    - Alice: **mere seconds**
    - Bob: **9 minutes**



\* Thanks to Danny Tabak of Aladdin for the implementation!



- **Physically sending a smartcard is OK once (or occasionally)**
  - Sending a smartcard every execution is unrealistic
- **However, this is not necessary**
  - Using secure messaging, Alice can write a new key  $k$  to the smartcard while keeping it secret from Bob

- **What else can be done in this model?**
- **Oblivious database search**
  - A client carries out a search on a database (retrieving a record via a keyword)
  - The server learns **nothing** about what the client searched for



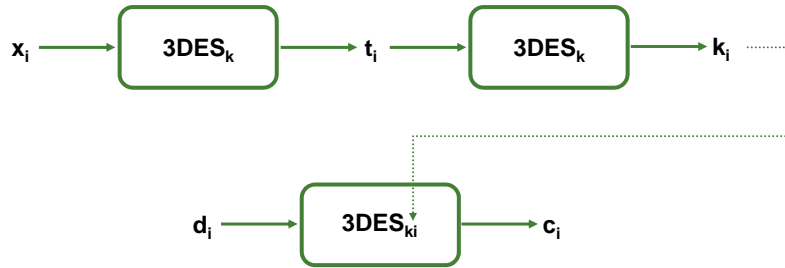
- **A trivial solution?**
  - The client downloads all of the database
- **Limiting information flow**
  - The aim of the solution is to **limit the amount of information** that the client obtains
  - The client is only allowed to carry out one search (or another predetermined number of searches)

- **How is it possible to limit the information flow without the server knowing what the client is searching?**
  - If the server knows, then it could just send the requested record
  - If the server doesn't know, how can we limit the number of searches the client carries out?

- **Classified databases**
  - One homeland security agency wishes to search for a suspect in a different agency's database
  - Allowing full access is dangerous
  - The identity of the suspect is also highly classified and so revealing it to the other agency is unacceptable

- **LexisNexis is a search engine for legal professionals**
  - Can search for case summaries etc.
- **There are a number of payment options: one of them is pay per search**
- **Such searches can be highly confidential**
  - An efficient solution to the above problem is highly desirable

- **Database structure**
  - Every record contains a keyword  $x$  (search attribute) and a record  $d$ 
    - The  $i^{\text{th}}$  record is denoted  $(x_i, d_i)$
  - The keyword  $x_i$  is unique in the database
- **Encrypting the database**
  - Compute  $t_i = 3DES_k(x_i)$  and  $k_i = 3DES_k(t_i)$ 
    - $t_i$  is the new keyword value and  $k_i$  is an encryption key
  - For every  $i$  encrypt  $c_i = 3DES_{k_i}(d_i)$



- **The server sends the client pairs  $(t_1, c_1), (t_2, c_2), \dots$** 
  - Recall that  $t_i = 3DES_k(x_i)$  and  $k_i = 3DES_k(t_i)$
- **The server sends a smartcard to the client with the key  $k$  inside**
  - The usage counter is set to the number of searches allowed to the client (times 2)
- **With keyword  $x$ , the client computes  $t = 3DES_k(x)$  using the smartcard**
  - If there exists an  $i$  for which  $t = t_i$ , then  $x$  is the  $i^{\text{th}}$  keyword
  - Compute  $k_i = 3DES_k(t_i)$
  - Decrypt  $c_i$  using  $k_i$  to obtain  $d_i$

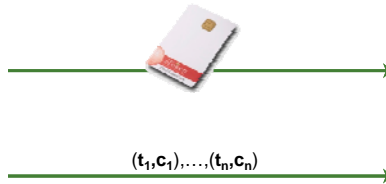
## The Protocol

### Server

Choose a random  $k$ ,  
initialize smartcard

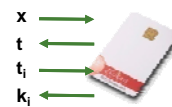
For every  $i = 1, \dots, n$  compute:

- $t_i = 3DES_k(x_i)$
- $k_i = 3DES_k(t_i)$
- $c_i = 3DES_{k_i}(d_i)$



### Client

Let  $x$  be keyword to search



Decrypt  $c_i$  using  $k_i$  to get  $d_i$

## Security Analysis

- **The server cannot learn anything**
  - It only sends information
- **The client learns only the predetermined number of queries**
  - Two smartcard operations are needed for obtaining a single  $k_i$
  - Without  $k_i$ , it is impossible to learn  $d_i$

- **How can we reuse the smartcard here to allow for many searches?**
- **A solution – background**
  - **Access-granted counter:** The 3DES computation can be limited to twice for every time a **test** is passed
  - The test can be a challenge/response using a strong cryptographic key

- **The server sends the encrypted database and smartcard to the client**
- **When the client wishes to carry out a search**
  - The client requests a **challenge** from the smartcard
  - The server provides the **response**
  - The client can then carry out one search (as required)

## The Full Protocol

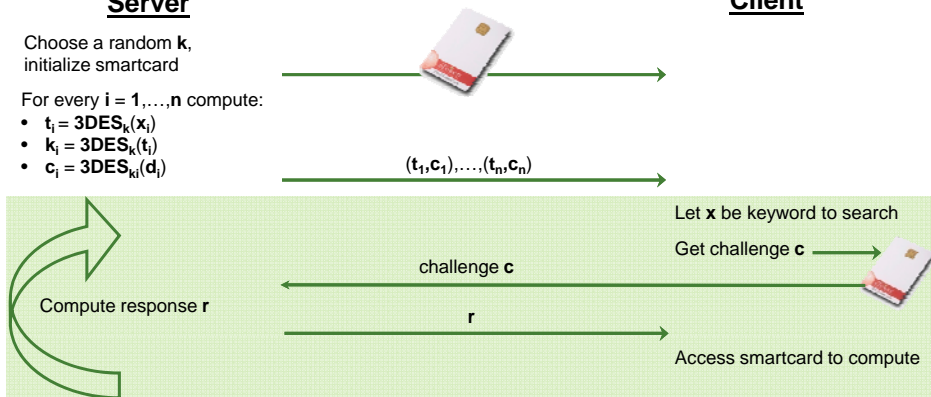
### Server

Choose a random  $k$ ,  
initialize smartcard

For every  $i = 1, \dots, n$  compute:

- $t_i = 3DES_k(x_i)$
- $k_i = 3DES_k(t_i)$
- $c_i = 3DES_{k_i}(d_i)$

### Client



## Proving Security

- **Small modifications have to be made to the protocol to achieve provable security**
  - Almost no effect on the efficiency
  - Follows same idea (but just need more keys in the smartcard)

- What about the more general case of document search by **keywords**?
- This can be solved using the previous solution, as follows:
  - Encrypt each document under a different key
  - For every keyword, define the “data” for this keyword to be the set of keys and document identifiers containing the keyword
  - Use the previous solution on this database

- Two people holding a value wish to check that they have the same value, without revealing anything else
- Applications
  - Two managers with the same confidential complaint
    - Are the complaints from the same employee?
  - Spies comparing knowledge of a secret code
  - Security agencies comparing the identity of a suspected mole
- Naïve solution doesn't work



- **Comparison is just a special case of set intersection**
  - Here, each set is of size **one**
- **Solution: use smartcard protocol**
  - Extraordinarily simple and efficient
- **Note: if inputs are **not equal**, nothing is revealed**

- **Assume that there are only 20 names to compare**
- **Phase one**
  - Line up 20 plastic cups with a name in front of each cup
  - Alice places a note in each cup: one of them says **YES** and the rest say **NO**
  - Bob does the same
- **Phase two**
  - Alice and Bob remove all of the names in front and shuffle all of the cups around
  - Alice and Bob check if there is one cup with two **YES** notes inside
- **Security...**



- **Privacy-preserving data mining**
  - Many data mining algorithms are far more complex than just set intersection, oblivious DB search and comparison
- **An important research challenge**
  - Come up with solutions to these problems that are
    - Secure, providing strong security guarantees
    - Efficient enough to be used in practice

- **Policy and technology are intertwined**
  - Technology needs to be built for problems that society needs solved and that law mandates
  - Privacy policy and law is driven by what is possible
- **Building policy around private data mining**
  - The technology exists – government can use it, and can provide higher privacy by mandating its use where necessary

- **Data sharing is crucial for the war on terror and many other government needs**
- **Privacy protections are important**
  - For our liberty
  - For preventing backlash that inhibits information flow
  - For cooperating with Europe and others who have different privacy regulations

- **Basic privacy**
  - Information must be protected from exposure
  - Information must be used appropriately
- **Privacy in a world of data sharing**
  - Computations on shared data must reveal **minimal** information



- **Smartcard-aided computation**
  - It is possible to construct secure protocols that are highly efficient
  - Efficient protocols do not yet exist for all tasks, but do for some
    - Set intersection and comparison
    - Oblivious database search
    - Oblivious document search

- **Promote higher awareness among policy-makers that it is possible to preserve privacy while mining **shared data****
- **Construct efficient protocols that meet the needs of policy and law**
- **Requires cooperation between**
  - Security experts and cryptographers
  - Legal, policy and privacy experts



Thank You



Andrew Lindell  
Private Data Mining and Citizens' Rights  
April 29, 2008



Legal Notice

© Copyright 2008 Aladdin Knowledge Systems Ltd. All rights reserved.

Aladdin, Aladdin Knowledge Systems, the Aladdin Knowledge Systems logo, eToken and eSafe are trademarks of Aladdin Knowledge Systems Ltd. covered by patents [www.aladdin.com/patents](http://www.aladdin.com/patents); other patents pending.

You may not copy, reproduce (or the like), or use in any other way whatsoever, whether directly or indirectly, any of the materials represented and/or disclosed herein without the express written consent of Aladdin.

Some of the information contained herein may be proprietary information of Aladdin or third parties and all text, images, graphics, trademarks, service marks, logos, trade names and other materials which are part of this communication are subject to intellectual property rights of Aladdin or third parties. The information herein is provided "as is" without any warranty, express or implied (by statute or otherwise), of any kind whatsoever. Aladdin does not undertake any obligation to update the information herein and it does not assume responsibility for errors or omissions.



Andrew Lindell  
Private Data Mining and Citizens' Rights  
April 29, 2008

