
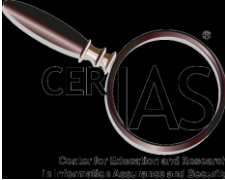


PURDUE
UNIVERSITY

CS62600: Advanced
Information Assurance


Privacy and Anonymity
8 September, 2009
Prof. Chris Clifton



What is Privacy?

Webster:
Freedom from unauthorized intrusion

- Intrusive
 - Is disclosure of the data not in the individual's best interest?





Intrusion



- Harm to individual
 - Physical, psychological, or perceived
 - How to measure?
- Use of data for other than approved purpose
 - Current standard in many areas
 - Too restrictive?
 - Too lenient?



Privacy



- “the ability to access and control one's personal information”
- Recognized by several treaties and protected by law
 - United States Healthcare Insurance Portability and Accountability (HIPAA)
 - The European Community Directive 95/46/EC
 - Privacy is about “*individually identifiable data*”



Terminology



- Private Data
 - Individually Identifiable
 - Sensitive
- Parties
 - Data subject
 - Person who the private data is about
 - Processor
 - Handles/manages private data
 - Recipient
 - Someone to whom data is disclosed
 - Adversary
 - One who would/could misuse private data



Regulatory Constraints: Privacy Rules



- Primarily national laws
 - European Union
 - US HIPAA rules (www.hipaadvisory.com)
 - Many others: (www.privacyexchange.org)
- Often control transborder use of data
- Focus on intent
 - Limited guidance on implementation



European Union Data Protection Directives



- Directive 95/46/EC
 - Passed European Parliament 24 October 1995
 - Goal is to ensure free flow of information
 - Must preserve privacy needs of member states
 - Effective October 1998
- Effect
 - Provides guidelines for member state legislation
 - Not directly enforceable
 - Forbids sharing data with states that don't protect privacy
 - Non-member state must provide adequate protection,
 - Sharing must be for "allowed use", or
 - Contracts ensure adequate protection
 - US "[Safe Harbor](#)" rules provide means of sharing (July 2000)
 - Adequate protection
 - But voluntary compliance
- Enforcement is happening
 - Microsoft under investigation for Passport ([May 2002](#))
 - Already fined by Spanish Authorities ([2001](#))



EU 95/46/EC: Meeting the Rules



- Personal data is any information that can be traced directly or indirectly to a specific person
- Use allowed if:
 - Unambiguous consent given
 - Required to perform contract with subject
 - Legally required
 - Necessary to protect vital interests of subject
 - In the public interest, or
 - Necessary for legitimate interests of processor and doesn't violate privacy



EU 95/46/EC: Meeting the Rules



- Some uses specifically proscribed
 - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life
- Must make data available to subject
 - Allowed to object to such use
 - Must give advance notice / right to refuse direct marketing use
- Limits use for automated decisions (e.g., creditworthiness)
 - Person can opt-out of automated decision making
 - Onus on processor to show use is legitimate and safeguards in place to protect person's interests
 - Logic involved in decisions must be available to affected person
- europa.eu.int/comm/internal_market/privacy/index_en.htm



US Health Insurance Portability and Accountability Act (HIPAA)



- Governs use of patient information
 - Goal is to protect the patient
 - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
 - A covered entity may not use or disclose protected health information, except as permitted or required...
 - To individual
 - For treatment (generally requires consent)
 - To public health / legal authorities
 - Use permitted where "there is no reasonable basis to believe that the information can be used to identify an individual"
- Safe Harbor Rules
 - Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
 - Name, location smaller than 3 digit postal code, dates finer than year, identifying numbers
 - Shown not to be sufficient (Sweeney)
 - Also not necessary
 - *Moral: Get Involved in the Regulatory Process!*



Contractual Limitations



- Web site privacy policies
 - “Contract” between browser and web site
 - Groups support voluntary enforcement
 - [TrustE](#) – requires that web site DISCLOSE policy on collection and use of personal information
 - [BBBOnline](#)
 - posting of an online privacy notice meeting rigorous privacy principles
 - completion of a comprehensive privacy assessment
 - monitoring and review by a trusted organization, and
 - participation in the programs consumer dispute resolution system
 - Unknown legal “teeth”
 - Example of customer information viewed as salable property in court!!!
 - [P3P](#): Supports browser checking of user-specific requirements
 - Internet Explorer 6 – disallow cookies if non-matching privacy policy
 - [PrivacyBird](#) – Internet Explorer plug-in from AT&T Research
- Corporate agreements
 - Stronger teeth/enforceability
 - But rarely protect the individual



Defining Privacy Modeling Real World



- What type of data the owner has?
 - Single table, relational, spatio-temporal, transactional, stream...
- What does the adversary know?
 - External public tables, phone books, names, ages, addresses...
- What is sensitive?
 - Medical history, salary, GPA...
- What is the RISK OF DISCLOSURE on both subject’s end and owner’s end?
 - Discrimination, public humiliation...
 - Court suits



Anonymization

- Goal: Not individually identifiable data
 - Specifically exempt from privacy laws
- Approaches
 - Remove identifiers
 - Generalization/suppression of non-identifiers
- Sensitive values still correct/usable
 - But what if generalized/suppressed values needed?




A Bogus Real World Model

- Data owner, hospital, has medical records
- Adversary knows names of the subjects
- Disease information is sensitive


Private Dataset

| Name | Age | Sex | Nation | Disease |
|-------|-----|-----|----------|---------|
| Obi | 17 | M | Turkey | Flu |
| Leta | 16 | F | Bulgaria | Flu |
| Padme | 23 | F | US | Obesity |
| Yoda | 25 | M | Canada | Tetanus |

Solution:
Remove
Unique Identifiers



Model Fails



← **Quasi Identifiers** →


| Age | Sex | Nation | Disease |
|-----|-----|----------|---------|
| 17 | M | Turkey | Flu |
| 16 | F | Bulgaria | Flu |
| 23 | F | US | Obesity |
| 25 | M | Canada | Tetanus |

Private Dataset


| Name | Age | Sex | Nation |
|-------|-----|-----|----------|
| Obi | 17 | M | Turkey |
| Leia | 16 | F | Bulgaria |
| Padme | 23 | F | US |
| Yoda | 25 | M | Canada |

Public Voters Dataset

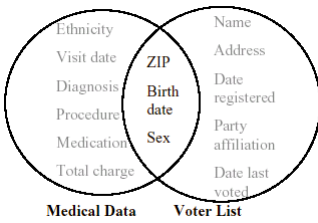
- In the real world, an adversary might have access to unique and **quasi identifiers** of the subjects
- In US, postal code, gender, birth date unique for 87%



Re-identifying “anonymous” data (Sweeney '01)



- 37 US states mandate collection of information
- She purchased the voter registration list for Cambridge Massachusetts
 - 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three



- Solution: k-anonymity
 - Any combination of values appears at least k times
- Developed systems that guarantee k-anonymity
 - Minimize distortion of results



Health Data Anonymization The Tradeoff



- Healthcare Data is Sensitive
 - Embarrassment
 - Economic
 - Legal
- Healthcare Data is Valuable
 - Research
 - Process optimization/improvement
 - Marketing



Anonymized Data



- HIPAA protects Individually Identifiable Health Information

| <i>Name</i> | <i>Addr.</i> | <i>Birth</i> | <i>Sex</i> | <i>Diagnosis</i> |
|-------------|--------------|--------------|------------|------------------|
| Alice | 47901 | 3/4/56 | F | ... |
| Bob | 47904 | 4/5/67 | M | ... |
| Chris | 47906 | 5/6/78 | M | Schizophrenic |



Anonymized Data

- HIPAA protects Individually Identifiable Health Information
 - Is this identifiable?
 - *Probably...*

| Name | Addr. | Birth | Sex | Diagnosis |
|------|-------|--------|-----|---------------|
| | 47901 | 3/4/56 | F | ... |
| | 47904 | 4/5/67 | M | ... |
| | 47906 | 5/6/78 | M | Schizophrenic |



HIPAA: De-Identifying Data

- A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable
 - Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
 - Documents the methods and results of the analysis that justify such determination
- The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:
 - Names, Location < 1st three digits of zip, dates < year, Tel/Fax/email/SSN/MRN/InsuranceID/Account/licence/VIN/License Plate Numbers, DeviceID, URL/IP, Biometric IDs, full-face photographs, any other unique identifiers; and
 - The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.



Anonymized Data

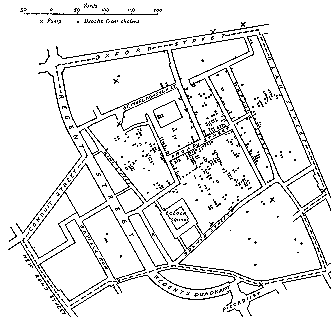
- HIPAA Safe-Harbor De-Identified Data
 - Is it useful?

| Name | Addr. | Birth | Sex | Diagnosis |
|------|-------|-------|-----|---------------|
| | 479xx | 56 | F | ... |
| | 479xx | 67 | M | ... |
| | 479xx | 78 | M | Schizophrenic |



Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
 - Is it useful?
- Dot chart by Dr. James Snow showing deaths from cholera in relation to the locations of public water pumps.
 - Observed that cholera occurred almost entirely among those who lived near (and drank from) the Broad Street water pump.





Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
 - Is it useful?
 - Is it enough?

| <i>Name</i> | <i>Addr.</i> | <i>Birth</i> | <i>Sex</i> | <i>Diagnosis</i> |
|-------------|--------------|--------------|------------|------------------|
| | 479xx | 56 | F | ... |
| | 479xx | 67 | M | ... |
| | 479xx | 78 | M | Schizophrenic |



Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
 - Is it useful?
 - Is it enough?

| <i>Name</i> | <i>Addr.</i> | <i>Birth</i> | <i>Sex</i> | <i>Diagnosis</i> |
|-------------|--------------|--------------|------------|-------------------------|
| | 479xx | 56 | F | ... |
| | 479xx | 67 | M | Uses Marijuana for Pain |
| | 479xx | 78 | M | Schizophrenic |



Anonymized Data

- HIPAA Safe-Harbor De-Identified Data
 - Is it useful?
 - Is it enough?

| Name | Addr. | Birth | Sex | Diagnosis |
|------|-------|-------|-----|---------------------------------|
| | 479xx | 56 | F | Uses Marijuana for Phantom Pain |
| | 479xx | 67 | M | Uses Marijuana for Pain |
| | 479xx | 78 | M | Schizophrenic |



Anonymization: Metrics

- K-anonymity (*Samarati, Sweeney*)
 - What is a good value of K?
 - EC95/46 just says “Identified”
 - US HIPAA safe harbor rules implies O(100)
 - Doesn’t protect sensitive data
- Discernibility (*Øhrn&Ohno-Machando*)
 - ℓ-diversity (*Machanavajjhala, Gehrke, Kifer, Venkatasubramaniam*)
 - Enforces distribution of sensitive information
- t-Closeness (*Li&Li*)
 - Enforces natural distribution of sensitive values



New Ideas



- **p -indistinguishability** (*Clifton, Kantarcioğlu, & Vaidya*)
 - Probability that a function exists that can distinguish individuals
- **(c, t) -isolation** (*Chawla, Dwork, Sherry, Smith & Wee*)
 - Does a point have fewer than t neighbors within distance c ?
- **δ -presence** (*Atzori, Nergiz, and Clifton*)
 - Control the probability that a given individual can be identified
- **Differential privacy** (*Dwork, McSherry, Nissim and Smith*)
 - Query results from a database with and without the individual should be indistinguishable



Data Obfuscation



- **Goal:** Hide the protected information
- **Approaches**
 - Randomly modify data
 - Swap values between records
 - Controlled modification of data to hide secrets
- **Problems**
 - Does it really protect the data?
 - Can we learn from the results?



Example: US Census Bureau Public Use Microdata



- US Census Bureau summarizes by census block
 - Minimum 300 people
 - Ranges rather than values
- For research, “complete” data provided for sample populations
 - Identifying information removed
 - Limitation of detail: geographic distinction, continuous → interval
 - Top/bottom coding (eliminate sparse/sensitive values)
 - Swap data values among similar individuals ([Moore '96](#))
 - Eliminates link between potential key and corresponding values
 - If individual determined, sensitive values likely incorrect
 - Preserves the privacy of the individuals, as no entity in the data contains actual values for any real individual.
 - Careful swapping preserves multivariate statistics
 - Rank-based: swap similar values (randomly chosen within max distance)
 - Preserves dependencies with (provably) high probability
 - Adversary can estimate sensitive values if individual identified
 - But data mining results enable this anyway!



Obfuscation



- Protect sensitive data
 - Recipient / processor doesn't see sensitive data
- Process: Add noise to data
 - Hides real sensitive values
 - Noise added by known (random) process
- Using the (noisy) data
 - Specialized techniques to remove impact of noise on aggregate data
 - Ex: Agrawal & Srikant SIGMOD'00



Quantification of Privacy

Agrawal and Aggarwal '01




- Intuition: A random variable distributed uniformly between $[0, 1]$ has half as much privacy as if it were in $[0, 2]$
- In general: If $f_B(x) = 2f_A(2x)$ then B offers half as much privacy as A
- Also: if a sequence of random variable A_n , $n=1, 2, \dots$ converges to random variable B, then privacy inherent in A_n should converge to the privacy inherent in B




Proposed metric



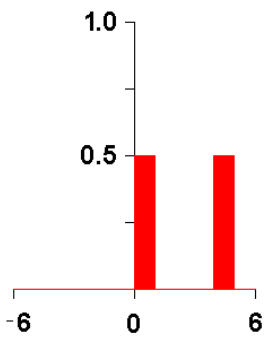
- Propose $\Pi(A) = 2^{h(A)}$ as measure of privacy for attribute A
- Uniform U between 0 and a : $\Pi(U) = 2^{\log_2(a)} = a$
- General random variable A, $\Pi(A)$ denote length of interval, over which a uniformly distributed random variable has equal uncertainty as A
- Ex: $\Pi(A) = 2$ means A has as much privacy as a random variable distributed uniformly in an interval of length 2

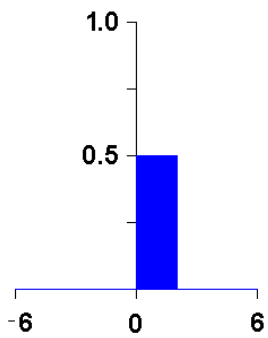


Example




- $f_X(x) = 0.5, 0 \leq x \leq 1$
- $f_X(x) = 0.5, 4 \leq x \leq 5$
- $f_X(x) = 0, \text{ otherwise}$






- Intuition from figures: X has as much privacy as a uniform variable over an interval of length 2 –
- Areas are the same:



Obfuscation: Issues



- How much is enough?
 - Dependent on adversary, sensitivity, individual?
 - Correlated values
- Is there a legal basis?
 - “Individually Identifiable Data” protected
 - Is wrong individually identifiable data different?



Restrictions on Results

- Use of Call Records for Fraud Detection vs. Marketing
 - FCC § 222(c)(1) restricted use of individually identifiable information
 - Until overturned by US Appeals Court
 - 222(d)(2) allows use for fraud detection
- Mortgage Redlining
 - Racial discrimination in home loans prohibited in US
 - Banks drew lines around high risk neighborhoods!!!
 - These were often minority neighborhoods
 - Result: Discrimination (redlining outlawed)
 - What about data mining that “singles out” minorities?



Regulatory Constraints: Use of Results

- Patchwork of Regulations
 - US Telecom (Fraud, not marketing)
 - Federal Communications Commission rules
 - Rooted in antitrust law
 - US Mortgage “redlining”
 - Financial regulations
 - Comes from civil rights legislation
- Evaluate on a per-project basis
 - Domain experts should know the rules
 - You’ll need the domain experts anyway – ask the right questions