PURDUE UNIVERSITY

CS42600: Computer Security

Data Privacy

Chris Clifton

18 April 2019

CERIAS

Center for Education and Research in Information Assurance and Security

---

PURDUE UNIVERSITY

What is Privacy?

- “The right to be let alone” - *Warren & Brandeis, 4 Harvard L.R. 193 (Dec. 15, 1890)*
  - My information protected so it doesn’t adversely affect me in the future
- Control over data
  - My information used only in ways I approve
- Issues:
  - Disclosure / sharing
  - Approved use
  - Recourse

2

PURDUE UNIVERSITY

Data Privacy: The Goal

- Protect the Individual
  - "Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified." – Charter of Fundamental Rights of the European Union
- Challenges: What do we mean by
  - "concerning" an individual
  - Protection
  - Consent
  - Access / rectified

European Commission

3

PURDUE UNIVERSITY

"Obvious" answers

- Concerning an individual
  - Has your name/address/other identifying information
- Protection
  - Only used/accessed in expected, intended, authorized ways
- Consent
  - You know and agree to what is done with the data
- Access/Rectify
  - You can see the data and correct errors

4

PURDUE UNIVERSITY

Consent?

[The Guardian](#)
Maev Kennedy
Thu 11 Jun 2009 07.17 EDT

American family's web photo ends up as Czech advertisement

Smiths from Missouri only heard about it when a friend travelling in Prague saw them on a grocery store poster

5

PURDUE UNIVERSITY

Could facebook have done this?

facebook didn't authorize it, it but could they?

Facebook Terms of … st, or upload
content that is cove … photos or videos)
on or in connection … -exclusive,
transferable, sub-li … icense to host, use,
distribute, modify, r … anslate, and create
derivative works of … acy and application
settings). This mea … oto on Facebook,
you give us permis … hers (again,
consistent with you … at support our
service or other Fa

Before 4/19/18, if shared with others, deleting your account didn't terminate these rights.

6

"Obvious" answers

- Concerning an individual
  - Has your name/address/other identifying information
- Protection
  - Only used/accessed in expected, intended, authorized ways
- Consent
  - You know and agree to what is done with the data
- Access/Rectify
  - You can see the data and correct errors

7

Concerning an Individual: IC 24-4.9-2-10 (Breach Disclosure)

IC 24-4.9-2-10 "Personal information"

Sec. 10. "Personal information" means:

(1) a Social Security number that is not encrypted or redacted; or

(2) an individual's first and last names, or first initial and last name, and one (1) or more of the following data elements that are not encrypted or redacted:

(A) A driver's license number.

(B) A state identification card number.

(C) A credit card number.

(D) A financial account number or debit card number in combination with a security code, password, or access code that would permit access to the person's account.

The term does not include information that is lawfully obtained from publicly available information or from federal, state, or local government records lawfully made available to the general public.

Other codes (e.g., spyware prohibition) have different definitions

8

PURDUE UNIVERSITY

The AOL Awakening

- In Aug 2006, AOL released its customers web searches for research studies
- 20 Million unique queries of 650K unique users
- <user-i
- NY Tim n individu... queries
  - Queries included “60 single men” “landscapers in Lilburn, Ga”
  - Many more queries contained enough information to uniquely identify the person
- *And it keeps going (Netflix, NYC Taxi, …)*

AOL fired its CTO over this issue;

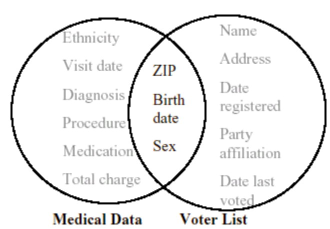
Two researchers were forced out

10

PURDUE UNIVERSITY

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

- 37 US states mandate collection of information
- Dr. Sweeney 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

Anonymity: The Goal

- Prevent Disclosure of Personal Information
  - GDPR: ‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly
  - Qatar Law 13 of 2016: Personal Data: Data belonging to an Individual with specified or reasonably specifiable identity whether through such Personal Data or through combining the same with any other data
  - *But still use the data where appropriate!*
- Problem: It can’t be done!
  - “Perfect” privacy requires zero utility (e.g., the data must be encrypted.)
  - As soon as we can use the data (e.g., decrypt), it is at risk

13

Why Perfect Privacy is Impossible *(Dwork, McSherry, Nissim, and Smith ‘06)*

- Background Knowledge
  - Adversary may already know a lot
  - Whatever we provide (even de-identified or anonymized data) may add to that knowledge
- It may just take that “last bit of knowledge” to give the adversary the ability to violate privacy
  - *We can formally prove 1 bit may be too much*
- The possibility is real
  - Garfinkel, Abowd, and Martindale, *Understanding Database Reconstruction Attacks on Public Data*, CACM 62(3): 3/19

14

PURDUE UNIVERSITY

What We Can Do

- Encryption
  - Reduce risk to minimal levels when data not in use
- Anonymization
  - Produce usable data that is hard to link to individuals
- Noise addition
  - Usable data where any link to individuals (or information we surmise about individuals) is guaranteed to be uncertain/suspect

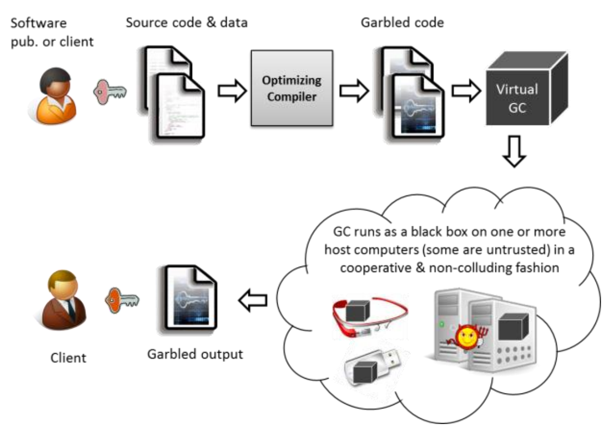
15

PURDUE UNIVERSITY

What We Can Do: Encryption

- Goal: Reduce risk to minimal levels when data not in use
- Encrypted Computation
  - Process the data while it is encrypted
  - Decrypt final output: Generalized, non-individual results
- Basic tools
  - Homomorphic Encryption, Commutative Encryption, Order Preserving Encryption
- Research Prototypes can accomplish many data processing and analysis tasks using these tools
  - Garbled Computing: Compute without revealing either the data or the program

- Garbled Computing.



17

PURDUE UNIVERSITY

What We Can Do: Anonymization

- Ensure protected/sensitive data not directly identifiable
  - Remove links between protected data and identifiers
- Generalize “quasi-identifiers”: Information that when combined with external data enables re-identification
  - Birth dates, addresses, workplace, etc.
  - E.g., instead of birth date, only give year
- Anonymized data still useful for data analysis
  - Goal is general knowledge, not learning specifics about individuals
- Example: “Anatomized” database from “Private Data in the Cloud” project

Patient	ID
Roan	1
Lisa	2
Roan	3
Elyse	4
Carl	5
Roan	6
Lisa	7
Roan	8

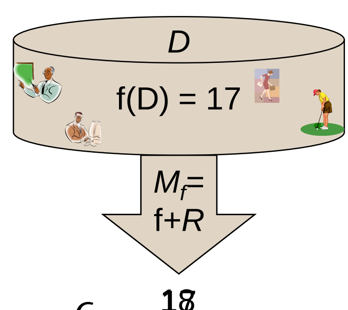
ID	Manufacturer	Drug Name
	Raphe Healthcare	Retinoic Acid
	Raphe Healthcare	Retinoic Acid
	Raphe Healthcare	Retinoic Acid
	Envie De Neuf	Mild Exfoliation
	Emedoutlet	Nexium
	Gep-Tek	Abiraterone
	Jai Radhe	Adapalene
	Hangzhou Btech	Cytarabine

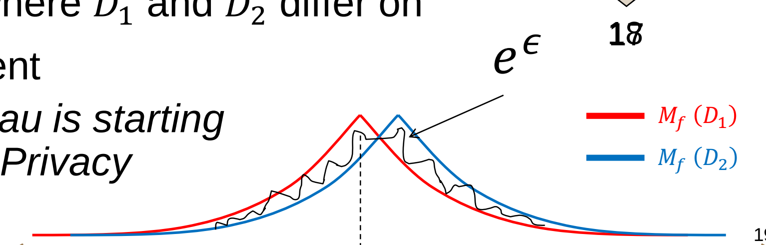
18

PURDUE UNIVERSITY

What We Can Do: Noise Addition

- Idea: Impact of noise on what we learn from the data larger than impact of any individual’s data
- Formally: For $S \subseteq Range(f)$, an ε-differentially private mechanism M satisfies $\frac{Pr[M_f(D_1) \in S]}{\Pr[M_f(D_2) \in S]} \leq e^{\epsilon}$ where D_1 and D_2 differ on at most one element
- *U.S. Census Bureau is starting to use Differential Privacy*





19

Achieving Differential Privacy

- Laplace Mechanism
  - Add Laplacian noise to the query result
  - Calibrate noise to the sensitivity of the query

$$Private\ f(D) = f(D) + Lap\left(\frac{\Delta f}{\epsilon}\right)$$

- Sensitivity
  - Captures the largest contribution to the result that can be made by one individual

$$\Delta f = \max_{D,D'} |f(D) - f(D')|$$

23

Another Example: Randomized Response *(Warner '65)*

- For each respondent with a yes/no value $f(D)$, flip a coin:
  - If heads, $Private\ f(D) = f(D)$
  - If tails, $Private\ f(D) = a\ second\ coin\ flip$
- True answer

$$\operatorname{avg} f(D) = \frac{1}{4}(1 - \operatorname{avg} Private\ f(D)) + \frac{3}{4} \operatorname{avg} Private\ f(D)$$

- Differentially private with $\epsilon = \ln 3$
  - Changing first coin flip changes epsilon

24

Exponential Mechanism

- The exponential mechanism $M_E(x, u, R)$ selects and outputs an element $r \in R$ with probability proportional to $\exp(\frac{\epsilon\, u(x,r)}{2\Delta u})$
  – x is database, u captures how much a given r distorts the outcome for the database x
  – Δu is sensitivity – maximum distortion r can cause across neighboring databases

25

Cool Properties of Differential Privacy

- Assume M is a differentially private mechanism
- Post-processing: $f \circ M$ is differentially private
  – Once the results are differentially private, anything we do with the results (that doesn't look back at the data) is still private
- Composition: M_1, M_2 are ϵ_1, ϵ_2-differentially private mechanisms
  – $M_1(x), M_2(x)$ is $(\epsilon_1+\epsilon_2)$-differentially private
  – If x, y disjoint, $M_1(x), M_2(y)$ is $\max(\epsilon_1, \epsilon_2)$-differentially private *Some caveats on this*

26

privacy parameter

- How to set ϵ
  - Open problem
  - No rule of thumb or guidelines
  - Physical meaning of ϵ
  - indistinguishable = unidentifiable ?

(a) large ϵ (b) small ϵ

27

Differential Identifiability *(Lee&Clifton, KDD'12)*

- Issue: What is the right value for ϵ?
  - Tells how far the *answer* is off
  - Want to bound probability of identification:
    $\Pr[i \in D \mid M_f(D)=R] \leq \rho$
- Differential Privacy easy enough to achieve
  - Adding Laplacian noise guarantees ε-differential privacy

$$M_f(X) = f(X) + Lap\left(\frac{\Delta f}{\epsilon}\right)$$

  - Somewhat more complicated for Differential Identifiability
    - But same basic approach/math

PURDUE UNIVERSITY

Identifiability

- Privacy game
  - $\mathcal{U} = \{u_1, u_2, \cdots, u_m\}$

Privacy mechanism

1. Pick a database $D \in \mathcal{U}^n$
   $D = (d_1, \cdots, d_n)$
   $D' = D - \{d_n\}$

3. $\mathcal{M}(D) = f(D) + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)$

4. Send $\left(\mathrm{D}', r = \mathcal{M}(D)\right)$

adversary

2. query f

5. Generate possible worlds
   $\psi_i = D' \cup \{u_i\}$

6. Guess who the nth individual is
   $\underset{i}{\operatorname{argmax}} \Pr[\mathcal{M}(\psi_i) = r]$

To limit adversary's confidence to ρ, what value of ϵ should we use?

29

PURDUE UNIVERSITY

Differential identifiability

- Practical definition
  - privacy based on differential privacy
  - probabilistic interpretation of individual identifiablity
- Definition
  - $\mathcal{M}$ satisfies ρ-differential identifiability if

$$\forall\, D' = D - \{i\}, \forall i \in U - D'$$

$$\mathbf{Pr}[\boldsymbol{I}(\boldsymbol{i}) \in \boldsymbol{I}_D | \, \boldsymbol{\mathcal{M}}_f(\boldsymbol{D}) = \boldsymbol{R}, \boldsymbol{D}'] \leq \boldsymbol{\rho}$$

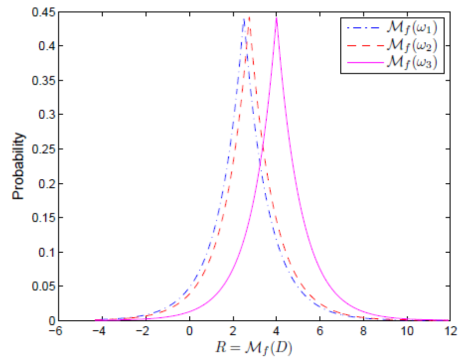
30

privacy parameter

- Simple example
  - U={1, 2, 3, 4, 5, 10}
  - D = {1, 2, 3, 10} , D'={1, 2, 3}
  - f=mean, R=5.4

ψ	f	$\epsilon = 1$	$\epsilon = 2$
$\omega_1 = \{1,2,3,4\}$	2.5	0.2353	0.1478
$\omega_2 = \{1,2,3,5\}$	2.75	0.2666	0.1898
$\omega_3 = \{1,2,3,10\}$	4	0.4981	0.6624

Table 1. $\Pr[\, D = \omega_i \mid R, D' \,]$



- DP-classfier [Cormode KDD 2011]
  - build an ϵ-DP naïve bayes classifier
  - can predict (potentially) sensitive information w.h.p.

Differential identifiability

- Assumption
  - uniform prior distribution
  - or, $t = \max_i \Pr[\psi_i = D]$ is known

- Relationship to DP

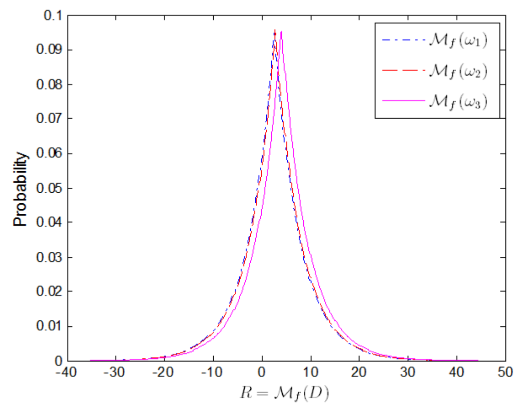
Any ϵ-differential private mechanism satisfies $\frac{1}{1+(m-1)e^{-\epsilon}}$-differential identifiability

  - inherits nice properties of DP
  - graceful degradation

PURDUE UNIVERSITY

revisiting the example

- Setting
  - ρ=0.4
  - $\lambda = \frac{S(f)}{\ln\frac{(m-1)\rho}{1-\rho}} = 5.21$
- Possible worlds
  - $\omega_1 = \{1,2,3,4\}$
  - $\omega_2 = \{1,2,3,5\}$
  - $\omega_3 = \{1,2,3,10\}$
  - $\Pr[\mathcal{M}_f(\omega_1) = R] = 0.0589$
  - $\Pr[\mathcal{M}_f(\omega_2) = R] = 0.0618$
  - $\Pr[\mathcal{M}_f(\omega_3) = R] = 0.0785$
  - $\boldsymbol{Pr[\omega_3 = D|R] = \frac{0.0785}{0.0589+0.0618+0.0785} = 0.3941}$



34

PURDUE UNIVERSITY

What We Can Do: Anonymization

- Ensure protected/sensitive data not directly identifiable
  - Remove links between protected data and identifiers
- Generalize “quasi-identifiers”: Information that when combined with external data enables re-identification
  - Birth dates, addresses, workplace, etc.
  - E.g., instead of birth date, only give year
- Anonymized data still useful for data analysis
  - Goal is general knowledge, not learning specifics about individuals
- Example: “Anatomized” database from “Private Data in the Cloud” project

Patient	ID
Roan	1
Lisa	2
Roan	3
Elyse	4
Carl	5
Roan	6
Lisa	7
Roan	8

ID	Manufacturer	Drug Name
	Raphe Healthcare	Retinoic Acid
	Raphe Healthcare	Retinoic Acid
	Raphe Healthcare	Retinoic Acid
	Envie De Neuf	Mild Exfoliation
	Emedoutlet	Nexium
	Gep-Tek	Abiraterone
	Jai Radhe	Adapalene
	Hangzhou Btech	Cytarabine

35

Problems with Anonymity

- Can we really prevent re-identification?
  - Experience says no
  - Big Data (*Variety*) makes it worse
- If we can, is the data still useful?
  - Currently having serious issues with anonymizing city-sized health information dataset

36

Myth: Anonymity is Broken

- Many academic papers with attacks on anonymization
  - E.g., deFinetti *(Kifer'09)*, Minimality *(Wong, Fu, Wang, Pei '07)*
  - Real-world failures (e.g., AOL)
- Reality: There is a risk
  - But risk may be acceptable (e.g., HIPAA safe-harbor rules do not eliminate risk of re-identification)
  - Differential Privacy provides provable limits on risk
  - **Any disclosure that provides utility also carries some privacy risk** *(Dwork'06)*

37

PURDUE UNIVERSITY

ℓ-Diversity

- Example using Bucketization
  - Anatomy (Xiao et al. (2006))
- k-Anonymous and ℓ-Diverse
- Every instance in IT can be matched to ℓ=2 instances in ST.

Age (A)	Zipcode (Z)	Job (J)	GID (G)
41	47905	Assoc. Prof	1
29	47906	Assist. Prof	1
30	47906	Assist. Prof	2
35	47907	Assoc. Prof	2
28	47906	Assist. Prof	3
47	47905	Prof.	3
45	47905	Prof.	4
31	47906	Assist. Prof	4

Identifier Table (IT)

ℓ =2

GID (G)	Income (I)
1	[100K-150K)
1	[50K-75K)
2	[75K-100K)
2	[50K-75K)
3	[75K-100K)
3	[100K-150K)
4	[100K-150K)
4	[75K-100K)

Possible Values

Sensitive Table (ST)

PURDUE UNIVERSITY

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.

PURDUE UNIVERSITY

Anonymized Data

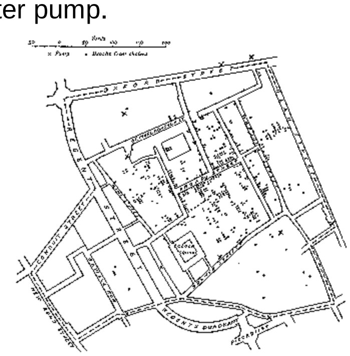
- HIPAA Safe-Harbor De-Identified Data
  - Is it useful?

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	479xx	56	F	…
	479xx	67	M	…
	479xx	78	M	Schizophrenic

PURDUE UNIVERSITY

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?

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?
  - Is it enough?

Name	*Addr.*	*Birth*	*Sex*	*Diagnosis*
	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

Myth: Anonymized Data Isn't Useful

- Differential Privacy: Noise added for privacy is often small relative to other sources of noise in the data
  - Can even improve statistical value of results *(Dwork et al. '17)*
- Machine Learning models learned from Anonymized Data can be as good or better than using the original data
  - Decision trees on k-anonymous data *(Iyengar'02)*
  - Support Vector Machines on anatomized data *(Mancuhan&Clifton'17)*
  - Nearest Neighbor on anatomized data

45

Machine Learning from Anonymized Data *(Mancuhan&Clifton'17)*

- Binary Classification task: *predict an attribute in IT given the other attributes in IT and the attribute in ST*
  - Example: predict age <35 or >=35 given job, zipcode and income
- What about predicting the attribute in ST table? (Example: income)
  - *Amounts to defeating privacy*
- Why do we care about using ST?
  - *Income may be useful to predict Job*

Age (A)	Zipcode (Z)	Job (J)	GID (G)
41	47905	Assoc. Prof	1
29	47906	Assist. Prof	1
30	47906	Assist. Prof	2
35	47907	Assoc. Prof	2
28	47906	Assist. Prof	3
47	47905	Prof.	3
45	47905	Prof.	4
31	47906	Assist. Prof	4

Identifier Table (IT)

GID (G)	Income (I)
1	[100K-150K)
1	[50K-75K)
2	[75K-100K)
2	[50K-75K)
3	[75K-100K)
3	[100K-150K)
4	[100K-150K)
4	[75K-100K)

Sensitive Table (ST)

Learning from Anonymized Data

- Anatomization: Possible to learn accurate models to classify data
- Can even outperform the models that are trained on original data in terms of
  - Error Rate (*K*-NN, Linear SV Classifier)
  - Convergence (1-NN)
- Can also reduce error compare to using attributes in IT alone
- Much better and practical than models for generalized/suppressed data
- Large training set helps…

PURDUE UNIVERSITY

Use Cases for Anonymity

- Public release
  - Challenging, given possible attacks on anonymity
- Protection from "accidental re-identification"
  - Ethical researchers could see their neighbor…
  - Model: HIPAA Limited Dataset
    - Easily re-identifiable, but only released under Data Use Agreement
- Reduce risk from data breach
  - Which would you rather have stolen, identifiable data or anonymized, *possibly* re-identifiable data
  - *Won't trigger many breach disclosure laws*

Can still obtain high quality analysis outcomes

48

PURDUE UNIVERSITY

What We Need: Legal Incentives

- "Notice and Consent" framework discourages application of technological advances
  - We can't guarantee your privacy, so please allow us to use your data in unsafe ways
  - U.S.: Enforcement action against Snapchat for promising to protect privacy and not doing a good enough job
    - Companies get away with not even trying, as long as they tell you so
- Can legal frameworks acknowledge that privacy is at risk?
  - Require efforts to manage, not eliminate, that risk

49

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

PURDUE UNIVERSITY

Fair Information Practices

1. Notice/Awareness
2. Choice/Consent
3. Access/Participation
4. Integrity/Security
5. Enforcement/Redress
   - Self-Regulation
   - Private Remedies
   - Government Enforcement

http://www.ftc.gov/reports/privacy3/fairinfo.shtm