

**Big Data's End  
Run Around  
Anonymity  
and Consent**

**By  
Solon Barocas  
And  
Helen Nissenbaum**

**Book Preview Conference  
New York • June 16, 2014**

# Big data

Computer Science

Information Science

Statistics and Analytics

Digital Technologies

## Paradigm

**Faith in the power of data, patterns,  
and models to know and decide**

# Privacy solutions?

## Anonymity & Informed Consent

- ` Anonymity severs link between ID and person, removes burden of protecting privacy
- ` Informed consent respects subject control and right to choose

# Anonymity & Consent Traditional Challenges

- ` How to anonymize effectively
- ` How to operationalize Informed Consent

# Key Theses of our article

- 1) Anonymity is not enough
- 2) Consent is irrelevant

# Anonymity: Challenges

Vanity searches [the AOL case]

Telling bio data [see Wilbanks]

Sparse sets [e.g. Census]

Triangulation [rich sets]

Re-ID attacks: db matching, linkage & differencing

[e.g. Narayanan+Shmatikov]

# Anonymity: “Solutions”

Aggregation

K-anonymity [Sweeney]

Differential privacy [Dwork, et. al.]

etc.

# The Real Value of Anonymity

- ` Reachability vs. namelessness
- ` Act and speak without fear of reprisal, ridicule
- ` Protective cloak for children, sick, needy
- ` Freedom from commercial reach
- ` Supports whistle blowing, peer review, voting
- ` Freedom to study, develop moral autonomy
- ` Enables zones free from answerability

It doesn't matter that ...

they don't know my name

“[i]f a company knows 100 data points about me in the digital environment, and that affects how that company treats me in the digital world, what's the difference if they know my name or not?”

## “Anonymous identifiers”

A website uses a formula to turn its users' email addresses into jumbled strings of numbers and letters. An advertiser does the same ... Both send their jumbled lists to a third company that looks for matches. When two match, the website can show an ad targeted to a specific person, but no real email addresses changed hands. (e.g. AdID)

## Inference

“We don't want the name. The name is noise.”

PII is not necessary for predictive modeling. Useful (“sensitive”) facts may be statistically inferred.



# Research underwritten by Anonymity

“[p]roactively discovering depressive symptoms from passive and unobtrusive Internet usage monitoring”

*Raghavendra et al., “Associating Internet Usage with Depressive Behavior Among College Students,” IEEE Tech. & Soc.31, no. 4 (Winter ‘12):73–80*

✓ Anonymity as namelessness

~~The value of anonymity~~

# Informed Consent

## Trouble with “notice & choice”?

- Hasn't worked: status quo
- Poorly understood (Turow, *et. al.*)
- Poorly implemented (“critical adherents”)
- Burdens individual with multiple unilateral contracts
- Burdens organizations with contriving reasonable, *gotcha*-proof privacy policies

# More trouble, with Big Data

## Indeterminate + Unending

*What actors? [Fitbit, NSA, Axiom, CVS, Udacity, PSE&G]*

*What information? [Given, inferred]*

*Under what constraints? [sell, barter, subpoena]*

## Unpredictable

*Actionable, unintuitive correlations*

*Impossible to foretell future findings*

# “Solutions”

- Operational challenges
- More digestible privacy policies
- More usable choice architectures
- More supple policies

But...

Transparency Paradox  
Comprehensible or Comprehensive?

# Public lives of others

## ❖ “With friends like these...”

- ❖ *Social networks: what friends reveal implicitly & explicitly*

## ❖ Tyranny of the Minority

- ❖ *Inference from representative sample*
- ❖ “multiple attributes can be inferred globally when as few as 20% of the users reveal their attribute information.”

Mislove et al., “You Are Who You Know: Inferring User Profiles in Online Social Networks.”



Consent is difficult?  
Consent is Irrelevant

## THEM

“Privacy and Big Data are incompatible!”

“Must make tradeoffs!”

## US

Don't confuse means with ends

Big Data is an opportunity to make things right

We propose

A Paradigm shift in Privacy Protection

*From procedure to substance*

# Lessons from the Biomedical Sciences

- Occurs within rich sets of standards & expectations
- Informed consent is given for a **Limited Waiver**
  - patient/subject interests
  - broad societal/public benefit

O'Neill and Manson (2012) *Redefining Informed Consent in Bioethics*

WASHINGTON  
SQUARE  
PARK



No  
bicycling



[www.nyc.gov/parks](http://www.nyc.gov/parks)

No  
skateboarding



[www.nyc.gov/parks](http://www.nyc.gov/parks)

**DO NOT  
FEED THE  
PIGEONS**





Standards & Expectations

Limited waivers



Standards and Expectations

Limited  
waivers

# Bring landscape into focus

- Contextualize consent
- Explicate risk and benefit and its distribution
- Explicate interest, rights, obligations
- Data operators must give good reasons in terms of interests and ethical, political, and contextual values