Reality Mining:
The End of Personal Privacy?

Anmol Madan, Ben Waber, Margaret Ding, Paul Kominers,
Dr. Alex (Sandy) Pentland

Human Dynamics Group, MIT Media Lab

Acknowledgements : Lanthe Chronis, Daniel Olguin Olguin

Overview

- Introduction
- 1st Example: Modeling Workplace Interactions Using Badges
- 2nd Example: Modeling Evolution of Opinions Using Mobile Phone Data
- Discussion: Legal Privacy Implications of our Work
- Caveat: We are not LAWYERS!

1st Example: Sociometric Badge for Workplace Interactions

- Infra-Red (IR) Transceiver
  - F2F Interaction
- 3-Axis Accelerometer
  - Movement, empathy…
- Microphone
  - Tone of voice, speaking speed…
- 2.4 GHz Radio
  - Proximity, location, …
- Bluetooth
  - Data transfer

1st Example: IT Firm

- Deployed badges at a Chicago data server configuration firm for one month
  - 30 participants
  - Create system specifications for customers
- Productivity metrics from company database
  - Job completion time
  - Job complexity
  - Errors…

1st Example: IT Firm

- Found that a one standard deviation increase in social cohesion increased performance by 10%
- Measure expertise by combining badge and task level data
- Predict 66% of the variance in productivity at the task level

Wu, Waber, Aral, Brynjolfsson, and Pentland, 2008
Waber and Pentland, 2009

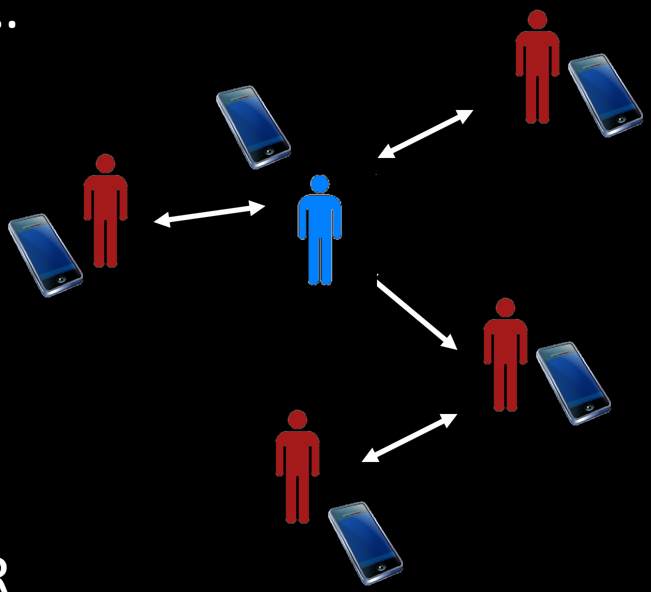
1st Example: Bank Call Center

- Studied Bank of America call center for one month
  - 80+ employees (4 teams)
  - E-mail, productivity, and survey data
- Social cohesion predicted productivity ($r = 0.61$)
  - The OPPOSITE of how call centers are managed!
  - Evidence that cohesion reduces stress as well
  - Reorganizing break structure in next experiment

Wu, Waber, Aral, Brynjolfsson, and Pentland, 2008
Waber and Pentland, 2009

2nd Example: Mobile data to model how 'things' spread in face-to-face networks

- Problem: Until now, real world face-to-face interactions were impossible to capture…
- Mobile phones provide:
  - Strength of ties
  - Entropy & homogeneity of behaviors
- Two aspects: adoption vs. causality
- Typical approach: threshold, cascade, SIR models with assumed mixing /exposure parameters



2^{nd} Example: Data Collection 1 Dorm, 1 Year

- MIT dorm, famous for tight-knit community + tech savvies, under the ‘microscope’
- 78 undergraduate participants for 1 academic year (started Fall 08)-- 80% of the dorm population *
- Used data collection mobile-phones as their primary phone, support 4 different operators, 6 different handsets
- Equivalent to 320,000 hours of data (~5 min scans)
  - 65,000 phone calls, 25,000 sms messages
  - 3.3 Million scanned bluetooth devices
  - 2.5 Million scanned 802.11 wlan APs

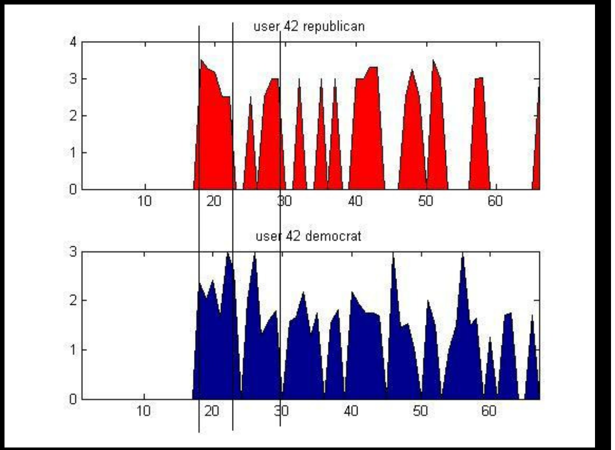
* with substantial help from Lanthe Chronis

2nd Example: Quantifying Exposure to Different Political Opinions

- Political Survey Responses (Likert scales)
  - Liberal or conservative (shifted n = 23)
  - Interested in politics (shifted n = 23)
  - Preferred Party (shifted n = 21)

- With Threshold / cascade / SIR-type epi models, key model parameter is exposure

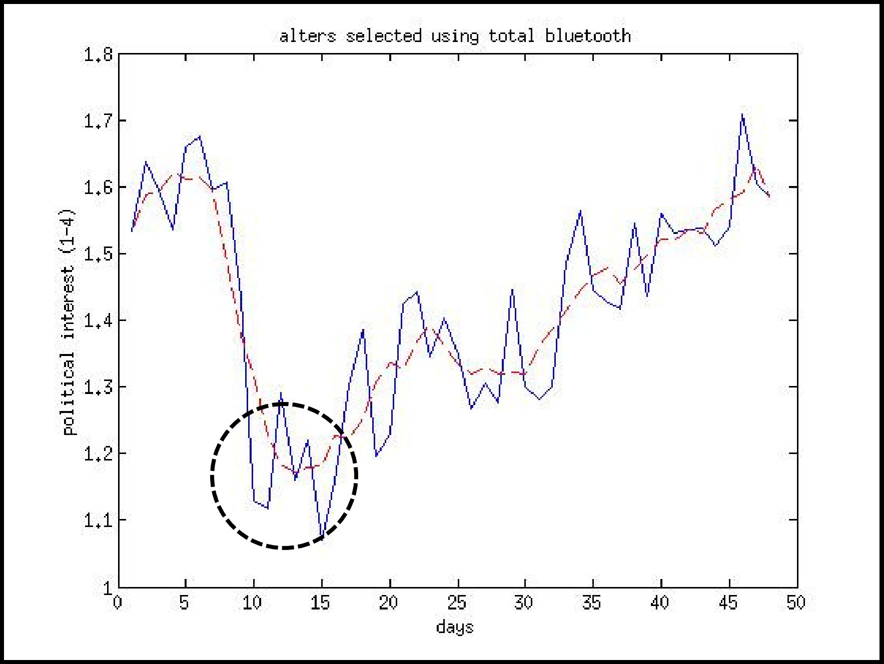
- Estimate daily exposure from mobile phone data:
  - Normalized i.e. what type of opinion is a person exposed to?
  - Cumulative i.e. to what magnitude of opinion A is a person exposed to?



Daily Republican & Democrat Exposure for one individual

2^{nd} Example: Loosely-Defined Homophily

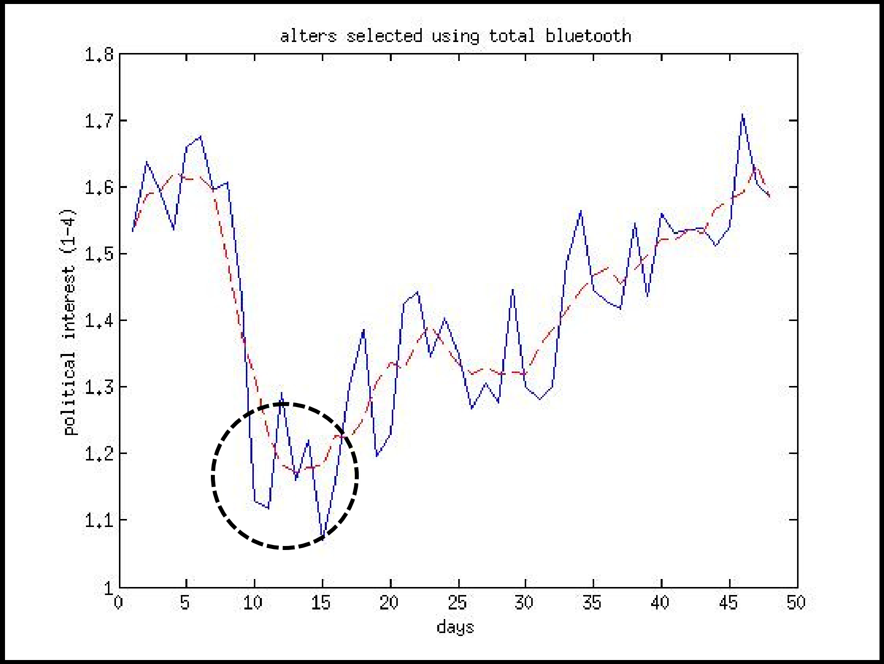
Averaged Difference Between an Individual's exposure and his/her political opinions, i.e. temporal convergence of opinions



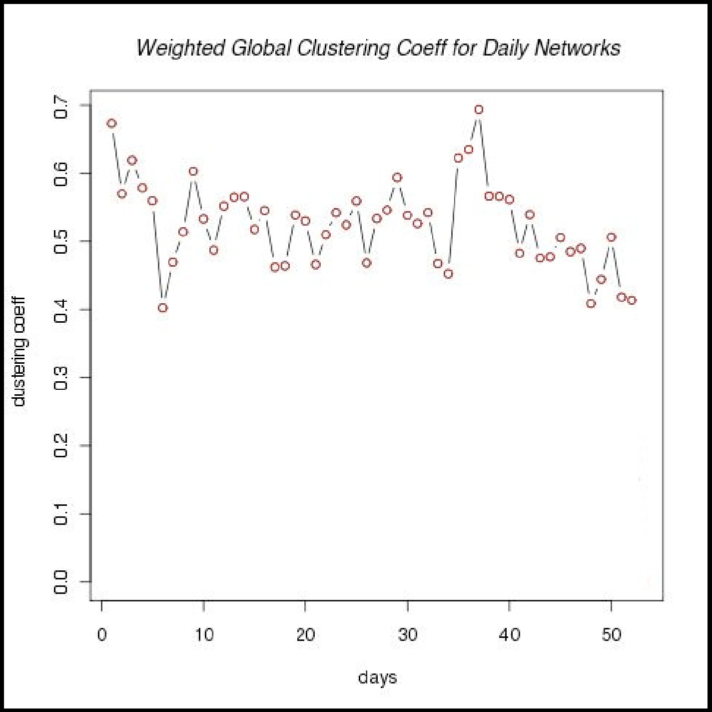
All residents
(day 0 = Oct 4^{th})

2nd Example: Loosely-Defined Homophily

Averaged Difference Between an Individual's exposure and his/her political opinions, i.e. temporal convergence of opinions



All residents
(day 0 = Oct 4th)

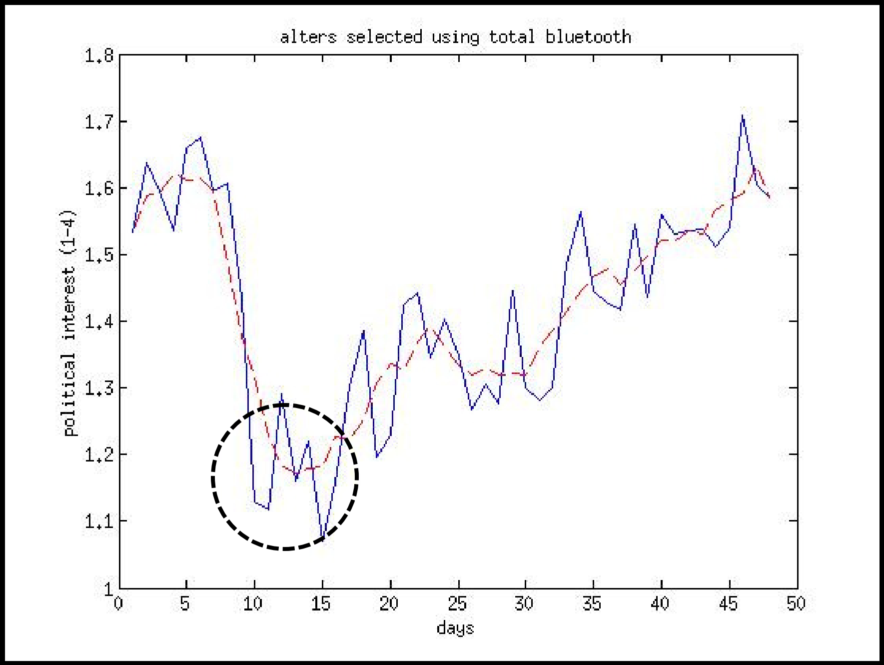


But the overall network structure remains invariant

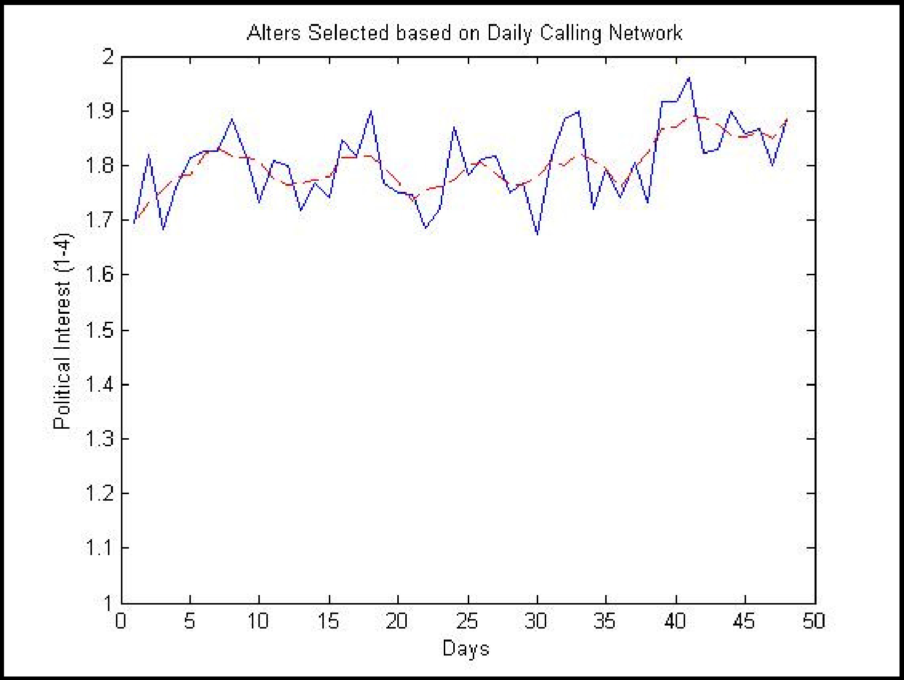
MIT Media Lab

2nd Example: Loosely-Defined Homophily

Averaged Difference Between an Individual's exposure and his/her political opinions, i.e. temporal convergence of opinions



All residents
(day 0 = Oct 4th)



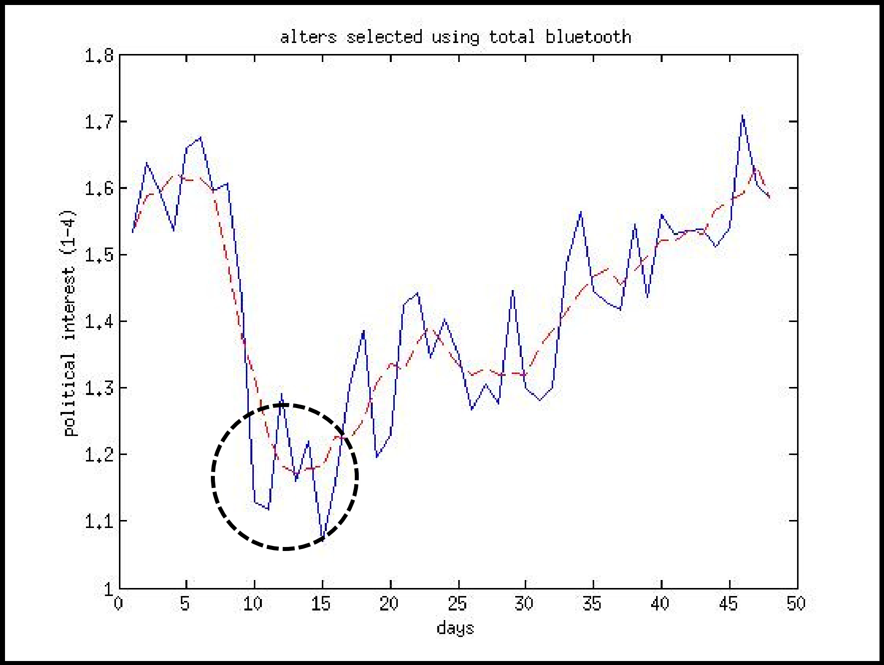
Phone calling network doesn't show the same structure that F2F interactions show

Human Dynamics Group

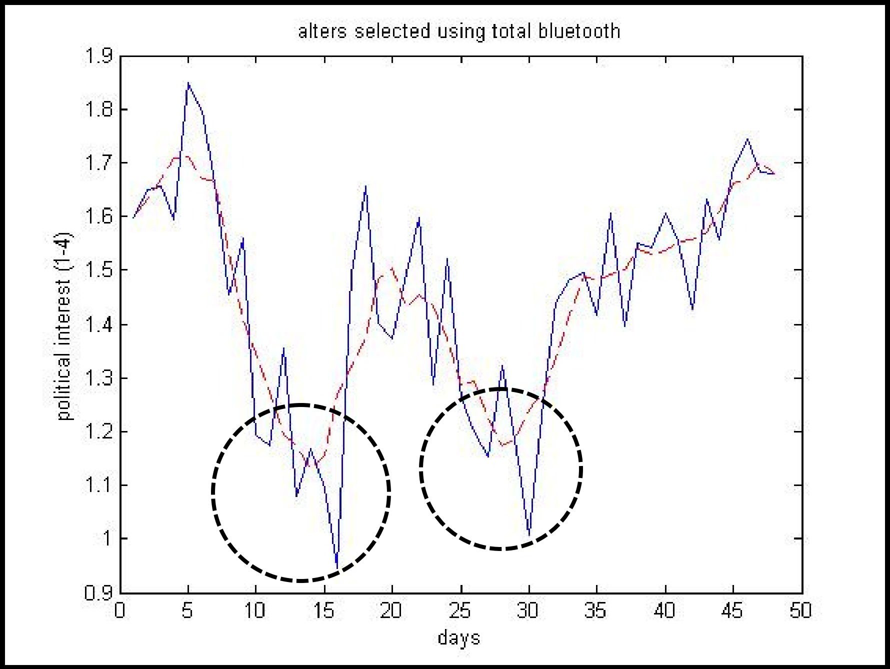
MIT Media Lab

2nd Example: Loosely-Defined Homophily

Averaged Difference Between an Individual's exposure and his/her political opinions, i.e. temporal convergence of opinions



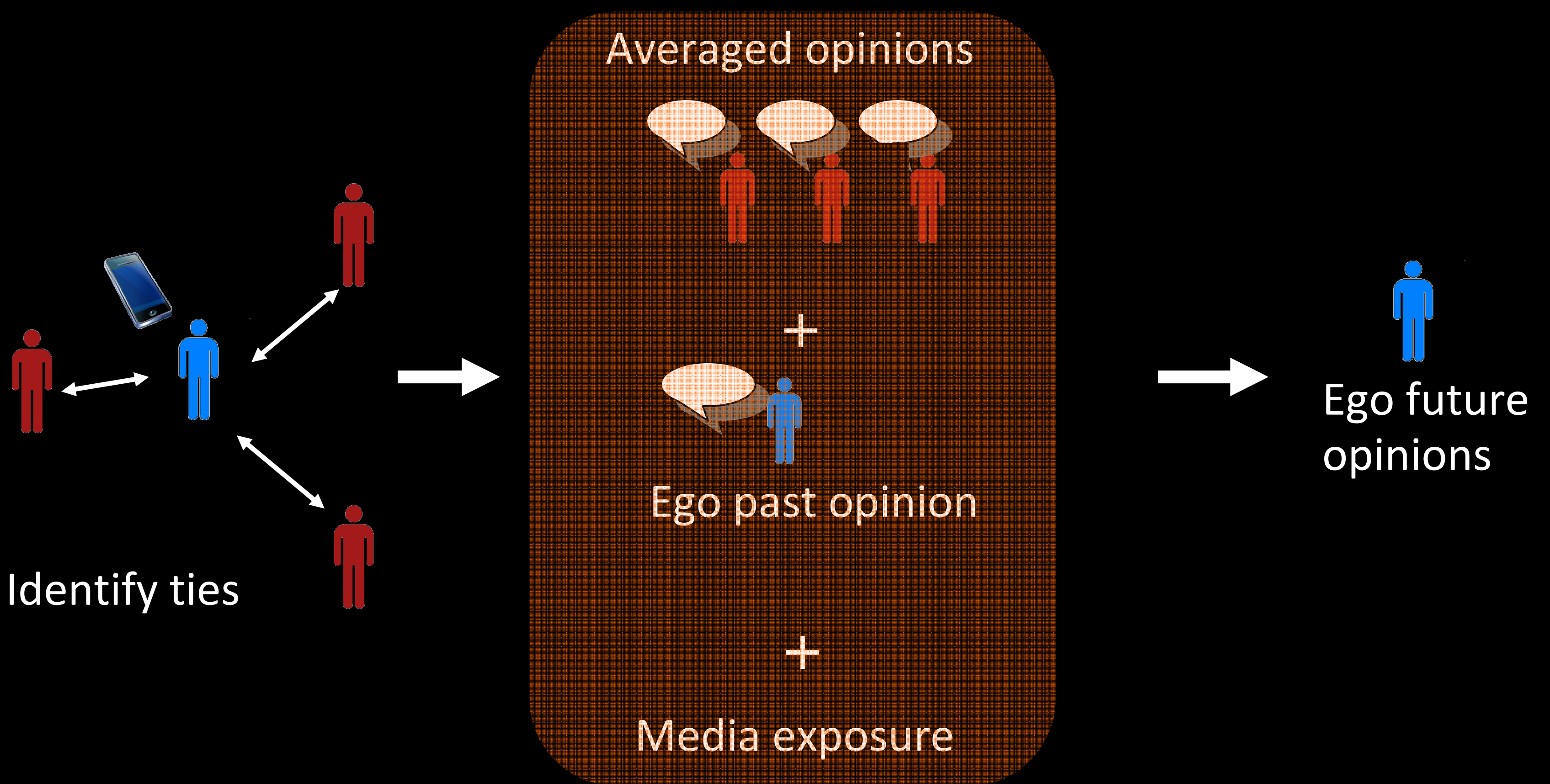
All residents
(day 0 = Oct 4th)



Freshmen Only
(day 0 = Oct 4th)

2nd Example: Likelihood of Adopting New Opinions

Opinion (x, T=t1) ~ Exposure (x, Δt), Opinions (x, t0), media (Δt)



2nd Example: Likelihood of Adopting New Opinions based on Estimated Exposure

- Ego's past opinion + friends' opinions are correlated with political opinions in Nov
  - political interest : R sqr = 0.75, p =< 0.0001
  - party preference : R sqr = 0.83, p < 0.0001
  - liberal or conservative : R sqr = 0.82, p < 0.0001

- Compare with just using ego's past opinion + control for media exposure– what is the value of 'automatically captured' ties and exposure?
  - political interest, party preference, liberal/conservative: 18%, 9% and 6% additional variance explained

- Stronger effects for freshmen:
  - political interest, party preference, liberal/conservative: 22%, 25% and 30% additional variance explained

Privacy Discussion

- Workplace Interactions: who owns employee data?
- Consumer Interactions: who owns end-user data?
- Data Anonymization: does it work?
- How are non-participants affected?

Real World Privacy (Quotes)

- "privacy aside, I personally have problems with people who don't live here leaving things in the dorm. Especially on a long-term basis, especially without permission, especially if they're trying to "study" us."

- "A quick poll of a cross section of the dorm" does not constitute permission. A significant fraction of xxx residents have a problem with this. Please do not place any devices in xxx"

- "What's the big deal? I've been recording all bluetooth activity from the ceilings of public spaces in the dorm for the past 9 years and posting all the data on xxxx. If you are concerned with who is recording your bluetooth devices, this is the perfect opportunity to change your privacy settings;

- "So, just because I do something in a lounge where people can see it doesn't make it legal for people to film me without permission and use it in a study. ... See also, the Fourth Amendment. "

Employee Privacy: Problem

- EU has more stringent data privacy policies than US
- In the US informing employees of monitoring makes data collection legal
  - Badge is analogous to unconcealed video surveillance
- Can this situation be improved?

Employee Privacy: Solution

- Third-party data collection and storage
- Employer would not have data ownership rights
- Aggregate statistics would be available
- Follow International Labor Organization guidelines
  - informed consent, equal access, secure storage, ‘employment-related reasons’

Consumer Data Ownership: End-users vs. Incumbent Service Providers

- Mobile Operators: strong laws enforced by FCC / Telecom Act around privacy of consumer data and non-disclosure to unrelated 3rd parties.
- Similar regulations apply for banks and financial institutions
- What happens when a consumer wants to ***force*** an MO to share data with a 3rd party (e.g. mint.com vs. BoA, SkyDeck vs. AT&T)?
  - Mobile operators ***required*** to share data
  - Banks and financial institutions ***permitted*** to share data

Data Anonymization

- Not secure in general, esp. for data about location and social-ties
- Recent attacks:
  - use embedded nodes to de-anonymize social network datasets
  - Use related auxiliary graphs to de-anonymize
- Use of anonymous data not legally specified. Possible alternatives: binning, resampling, aggregate stats

Impact on Non-participants

- Real-world applications: non-participants are likely affected
- Two interpretations:
  - ethical / IRB : stronger, protects non-participants
  - Legal : murky
- Example:
  - if a non-participant is broadcasting BTIDs, will be automatically captured by the system
  - there may be no legal expectation of privacy with this data (reference Smith vs. Maryland, for call logs)
  - No contractual agreement between app developer and non-participant

Summary

- Illustrated how we can model human behavior – both workplace and for end-users, using badges and mobile phones
- Data ownership in the workplace: recommend International Labor guidelines, fair rights to employees, third-party participation
- Data ownership for consumers: Should be able to *use their own data*, even if collected by service providers
- Anonymization: removing personal identifiers doesn't ensure privacy
- Impact on Non-participants: complex question for real-world apps