Privacy invasion or innovative science?

Conrad Lee

University College Dublin

December 28, 2011

Summary

Privacy issues are preventing a leap forward in study of human behavior
By preventing the collection and public dissemination of high-quality datasets.

Are academics overly-sensitive to privacy issues?
I think so. In many cases, your privacy is already an illusion.

Summary

Privacy issues are preventing a leap forward in study of human behavior
By preventing the collection and public dissemination of high-quality datasets.

Are academics overly-sensitive to privacy issues?
I think so. In many cases, your privacy is already an illusion.

Privacy and social network research

- From the ‘practice oriented’ perspective of a PhD student
  - Disclaimer: I don’t specialize in privacy issues

Outline

Outline

Digital footprints: revolution for the social sciences

We're on the cusp of a new way of doing social science... Our predecessors could only dream of the kind of data we now have (Nicholas Christakis)

- Some questions have remained stubbornly unanswered:
  - Do beliefs/taste determine friendship, or vice versa?
  - Opinion leaders?
  - Is obesity contagious?
- Data from social media sites (and communication services) solve a few major problems:
  - include interaction network
  - observed ("revealed") rather than survey
  - large-scale
  - longitudinal

Digital footprints: revolution for the social sciences

We're on the cusp of a new way of doing social science... Our predecessors could only dream of the kind of data we now have (Nicholas Christakis)

- Some questions have remained stubbornly unanswered:
  - Do beliefs/taste determine friendship, or vice versa?
  - Opinion leaders?
  - Is obesity contagious?
- Data from social media sites (and communication services) solve a few major problems:
  - include interaction network
  - observed ("revealed") rather than survey
  - large-scale
  - longitudinal

Digital footprints: revolution for the social sciences

We're on the cusp of a new way of doing social science... Our predecessors could only dream of the kind of data we now have (Nicholas Christakis)

- Some questions have remained stubbornly unanswered:
  - Do beliefs/taste determine friendship, or vice versa?
  - Opinion leaders?
  - Is obesity contagious?
- Data from social media sites (and communication services) solve a few major problems:
  - include interaction network
  - observed ("revealed") rather than survey
  - large-scale
  - longitudinal

So, where's the data?

- Data re-use still the exception rather than the norm, leading to
  - Problems of replicability (e.g., obesity contagion)
  - Hard to build incrementally
  - Inefficiency
- Obstacles:
  - Ethics
  - Cooperation/threat of service providers
- Specialized or deficient datasets are shared
  - lacking gender, age, socio-economic, ethnicity
  - missing much of user's social universe
  - Ideally an isolated village of smartphone users

Outline

Outline

Tastes, Ties & Time

- A dataset sufficient for studying diffusion:
  - A relatively self-contained social group (students)
  - A service used heavily by all members of that group (facebook)
  - Resources to manually annotate the data (NSF funding)
- Includes information on
  - Favorite books, music, films
  - Gender, Socio-economic, race, academic major
- An unprecedented dataset (Lewis, 2008 [3])

Tastes, Ties & Time

- Data collected from small university in New England over four years
- Data must be made public (requirement of NSF funding)
- Ethical aspects approved by Harvard IRB
- facebook approved

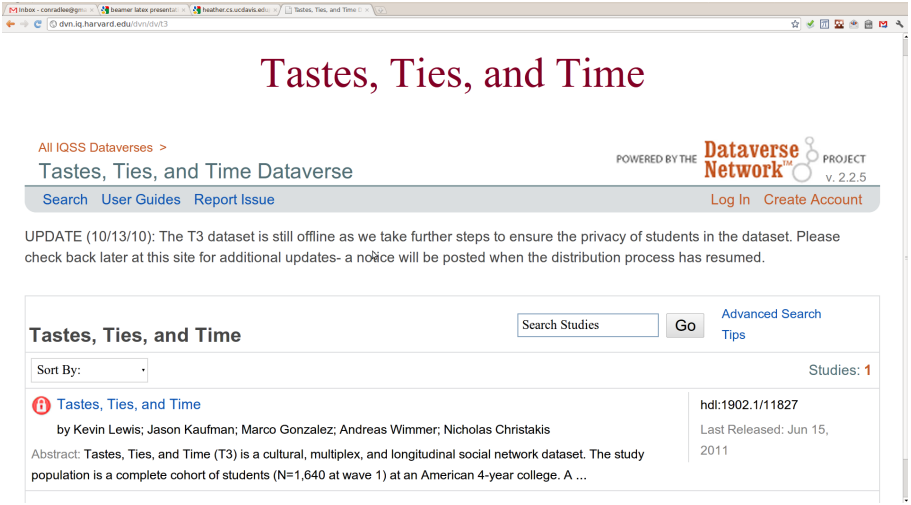
Privacy: good faith effort, but incompetent effort

And some serious ethical problems

- Measures taken
  - Names, contact info removed
  - Many attributes encoded
- Anonymity of dataset was quickly and easily cracked
- Data from Harvard class of 2009
- Serous criticisms of ethics
  - No consent or even notification
  - No way to opt out (asking would “frighten people unnecessarily”)
  - Profiles scraped by privileged students in same college network
  - Scrapers (embedded students) have special access to private data

Tastes, Ties & Time

Dataset quickly taken down, not currently publicly distributed
Still used by Harvard researchers



Outline

Facebook 100

The evil cousin of *Tastes, Ties, Time*

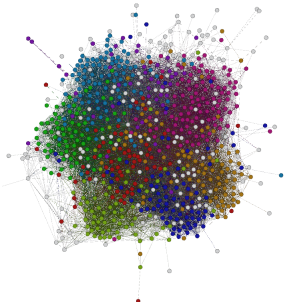


Figure: Caltech network visulized in Gephi

- Appeared in early 2011 [5]
- Data from September, 2005 (Facebook5 from June, 2005)
- Directly from facebook (from Adam D'Angelo, CTO)

Facebook 100

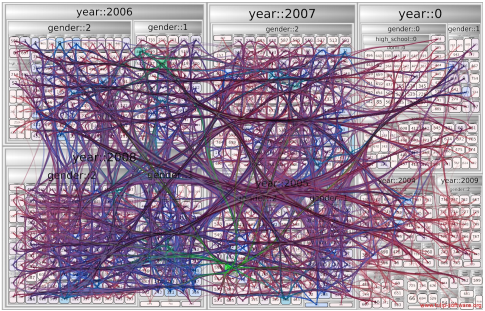


Figure: Caltech network visulized in Tulip

- For 100 U.S. universities, this dataset contains:
  - complete friendship network
  - attribute data (where available) on
    - gender
    - dorm
    - academic major
    - high school

Privacy of 1.2 million facebook users compromised?

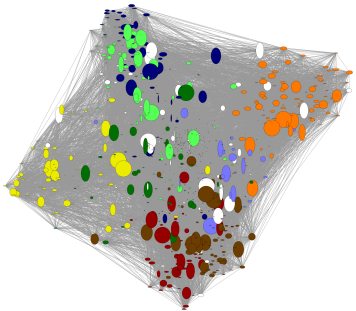


Figure: Caltech network visulized in Visone

- Friendship (and attribute?) data regardless of privacy settings
- But
  - Names removed
  - Attribute values encoded

Can the Facebook100 be cracked?

- Yes. See [1]
- But that requires me to have an exact subgraph from earlier.
  - Could I identify myself?
  - In half an hour, with high probability, narrowed myself down to one of 15 profiles

But does the Facebook100 need to be cracked?

- No. See
  http://michaelzimmer.org/2011/02/15/facebook-data-of-1-2-million-users-from-2005-released/
- Data was released with original facebook ids.
- This appears to be a mistake - data was taken down
- On bittorrent, for parser, Google “Facebook100 parser”

Finding myself

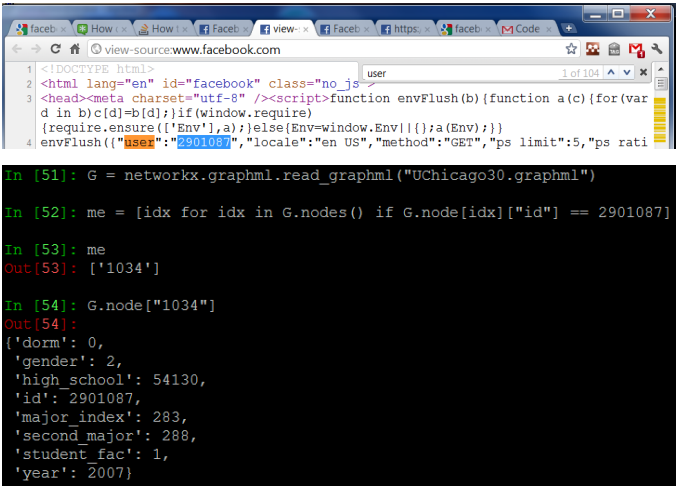


Figure: Uncovering personal data in Facebook100 dataset is easy

Finding Zuckerberg

```
In [20]: G = nx.graphml.read_graphml("Harvard1.graphml")

In [21]: zuckerberg_anon_id = [n for n in G.nodes() if G.node[n]["id"]==4][0]

In [22]: G.degree(zuckerberg_anon_id)
Out[22]: 156

In [23]: len([n for n in G.nodes() if G.degree(n) > 156])
Out[23]: 4055
```

Figure: Uncovering personal data in the Facebook100 dataset is easy

Outline

What about Tastes, Ties, Time?

- Data pulled to protect privacy of users
- In the meantime, Facebook releases data anyway
- Users' privacy is already (quietly) compromised
- Why not distribute the Tastes, Ties, Time dataset?

Conceptions of privacy

- Two privacy paradigms [6]
  - ***Harm based***
    - If hackers or others wishing to do harm don't get the data, everything is fine
    - Academics uninterested in identities can ethically use facebook data
  - ***Dignity based***
    - Concerns arise even if no harm takes place
    - If data stripped out of intended sphere, then basic human dignity of user has been compromised
- Effective research environments adopt the harm-based paradigm

Conceptions of privacy

- Two privacy paradigms [6]
  - ***Harm based***
    - If hackers or others wishing to do harm don't get the data, everything is fine
    - Academics uninterested in identities can ethically use facebook data
  - ***Dignity based***
    - Concerns arise even if no harm takes place
    - If data stripped out of intended sphere, then basic human dignity of user has been compromised
- Effective research environments adopt the harm-based paradigm

A questionable logic

The currently accepted policy

- You can exploit sensitive data for your own academic research (e.g. T3, Facebook100)
  - Just don't share it
- Ostensible explanation:
  - academic use is allowed, because academics do no harm
  - if we don't share it, it won't be used maliciously
  - ...because malicious users can't collect this data themselves?...

A questionable logic

The currently accepted policy

- You can exploit sensitive data for your own academic research (e.g. T3, Facebook100)
  - Just don't share it
- Ostensible explanation:
  - academic use is allowed, because academics do no harm
  - if we don't share it, it won't be used maliciously
  - ...because malicious users can't collect this data themselves?...

Such data leaks are the norm

- It's hard to maintain privacy and accessibility simultaneously
  - approx 75% of fb users left profile visible to "networks" (Jernigan et al, 2009)
  - These profiles visible by avg. 102,000 users
  - StudiVZ was notoriously insecure
  - Pete Warden's apparently legal facebook collection (210 million profiles)
  - Large twitter, foursquare datasets
- And these are just the ones we've heard about...
  - exploits of malicious users
  - exploits of big brother

Why does this implicit policy exist?

Who benefits from this policy?

- **Not users**, who are less aware of vulnerabilities
  - even though malicious parties may silently be exploiting them
- **Not science**, which is held back by
  - lack of high-quality datasets
  - lack of replicability
  - even though intentions are not malicious
- **Service providers benefit** (e.g. Facebook)
  - avoid bad press
  - avoid lawsuits
- **Malicious users benefit**
  - vulnerabilities remain unknown
  - confident users share more sensitive information

Why does this implicit policy exist?

- Researchers fear the wrath of service providers such as facebook
  - Ostensibly data not shared for privacy concerns (prevent malicious use)
  - However, those malicious entities likely have access to this data already, and perhaps more
  - Suggests a true motivation: academia fears the wrath of service providers like facebook
  - With good reason: the case of Pete Warden

Three cases

- Data can be collected:
  - publicly, without any agreement
  - publicly, with agreement to not distribute (e.g., through API)
  - privately to researchers, with agreement to not distribute
- If service providers leak out data easily, then why should academia not share datasets?
- Do service providers attempt to maintain privacy through the threat of lawsuits?
  - Is such a policy effective only for preventing research?
  - Are only those prosecuted who make weaknesses public?

Tastes, Ties, & Time

Concrete cases are grey

- Tastes, Ties, and Time is a grey area: could malicious individuals have collected this data?
  - In reality, yes. Ironically, the data is already released in the Facebook100 (and perhaps elsewhere)
  - Furthermore, anyone with a Harvard account could have collected much of the data if it hadn't already been released

Outline

“Enhancing” datasets

- Without well-curated datasets, researchers might get creative
- Social graph useful for inferring user attributes
- Need attribute values for only 20% to infer rest with 80% accuracy (Mislove, 2010 [4])
- Using logistic regression, Jernigan & Mistree, 2009 [2] were with high accuracy able to identify gay men
- How far should academia push this research? Should we enhance our own datasets with it?

Bibliography I

L. Backstrom, C. Dwork, and J. Kleinberg.
Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography.
In *Proceedings of the 16th international conference on World Wide Web*, pages 181–190. ACM, 2007.

C. Jernigan and B.F.T. Mistree.
Gaydar: Facebook friendships expose sexual orientation.
First Monday, 14(10), 2009.

K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis.
Tastes, ties, and time: A new social network dataset using facebook. com.
Social Networks, 30(4):330–342, 2008.

Bibliography II

A. Mislove, B. Viswanath, K.P. Gummadi, and P. Druschel.
You are who you know: Inferring user profiles in online social networks.
In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

Amanda L. Traud, Peter J. Mucha, and Mason A. Porter.
Social structure of facebook networks.
CoRR, abs/1102.2166, 2011.

M. Zimmer.
But the data is already public: on the ethics of research in facebook.
Ethics and information technology, 12(4):313–325, 2010.