Response to talks at Big Data and Privacy in Human Subject Research (1st day)

Hendrik Drachsler, @hdrachsler
Welten Institute Research Centre, Open University of the Netherlands
Presentation given at: NSF expert meeting on ‘Big Data and Privacy in Human Subjects Research’ (#BDEDU) 11 November 2014
WhoAmI

• Hendrik Drachsler, Open University of the Netherlands

• Research topics:
  Personalization,
  Recommender Systems,
  Learning Analytics,
  Mobile devices

• Application domains:
  Science 2.0
  Health 2.0
Research communities

dataTEL
Data-driven Research and Learning Analytics
EATEL-Special Interest Group

Hendrik Drachsler (a), Katrien Verbert (b)

(a) CELSTEC, Open University of the Netherlands
(b) Dept. Computer Science, K.U.Leuven, Belgium
Who are the good guys?
They brought together some Super Hero’s

Who of you considers him- herself to be a Super Hero?

- You are passionate about what you are doing.
- You shape the future of society.
- You touch ethical questions with your super power.
- You want to follow societal norms and advance those.

With Big Power comes great responsibilities.
Big Power -> Big Data = Repurposing data

Jawbone data repurposed to measure earthquake strength
Big Data is the new truth (the ultimate truth?)

LAK Challenge – the many faces of a small dataset

Analysis

Exploration & Visualisation

Search & Recommendation

Correlation & Enrichment
Big Data is the new truth (the ultimate truth?)

Inaccurate Google Flue trend measures compared to CDC
Big Data has the potential to change Education

- First time monitoring learning while it happens
- Personalize Education
- Identify students at Risk
- Learning Measures on demand
- More ...
Some Questions, Super Hero’s

Some Demographics: Who of you are data scientists, legal or educational experts?

Who of you read TOC of your online services?

Who of you cares about his/her privacy?

Who sees Privacy and Legal regulations as a burden we need to overcome?
Learning Analytics Research Issues

Learning Analytics research always raises the P-Word in EU (University of Amsterdam, 2014)

This stops innovation and advancing research (dataTEL 2010)
What is the problem with Privacy?

- Privacy changes over time
- Privacy is bind to context
- Privacy is bind to culture

Slide supported by Tore Hoel, @Tore
What if I would know ...

• How many days you have NOT been at school without any excuse.
• All read and written pages, and what your annotations have been.
• The people you hangout with in your youth.
• If you cheated in a test and how many attempts you needed for your math class.

• What if I use all those information and predict your chances to be good or bad in a certain job after school?
• How representative and reliable is this data I’m capturing to predict those chances?

• And what if all this information will be last forever!
Approaches to prevent another inBloom ...

• Transparency (Purpose of analysis, Raw data access, opt-out)
• Data Security
• Contextual Integrity (Smart Informed Consents)
• Anonymisation & Data degradation
Ethics & Privacy Issue in the Application of Learning Analytics (#EP4LA)

Hendrik Drachsler, @hdrachsler
Welten Institute Research Centre, Open University of the Netherlands
Presentation given at: NSF expert meeting on ‘Big Data and Privacy in Human Subjects Research’ (#BDEDU) 11 November 2014
Who we are

LACE Network

LACE Consortium

LTCS Group

LACE Network

Universities and Institutes:
- Deusto
- MELIUS
- EDUCAUSE
- Nova Scotia
- Mines-Télécom

Organizations:
- Open Knowledge Foundation
- GSIC-EMIC
- IMS Global Learning Consortium
- Manchester Metropolitan University
- University of Liverpool
- Skolverket
- Tamura Laboratory

Groups:
- LTCS Group
- Kennisnet

Academia:
- University of Exeter
- University of Brighton
- Anglia Ruskin University

Other:
- SURF
- Wien
- Lunds Universitet

Date and Author:
FP7 LACE – Hendrik Drachsler, @Hdrachsler, 28 October 2014
Building bridges between research, policy and practice to realise the potential of learning analytics in EU
Data Geology

- Are our instruments measuring what we expect them to measure?
- Can we isolate the noise in the data?
- Are the measures accurate?
Evidence.laceproject.eu
Privacy as Showstopper – The inBloom case

InBloom Student Data Repository to Close

By NATASHA SINGER APRIL 21, 2014 1:21 PM 94 Comments

In 2013 the database held information on millions of children.

Privacy as Showstopper – The inBloom case

New York State is planning to share your child’s confidential information with private corporations

New York State has agreed to share confidential student and teacher data with a Gates-funded corporation called inBloom Inc. All this is happening without parental consent:

1. Confidential data will include your child’s personally identifiable information, including name, address, grades, test scores, detailed disciplinary and health records, race, ethnicity, economic status, disabilities & other highly sensitive data.

2. Information will be collected into a “data store” with an operating system built by Wireless Generation, a subsidiary of News Corporation. News Corp. is owned by Rupert Murdoch & has been found to have illegally violated privacy in Great Britain & in the U.S.

3. The data is being stored on a vulnerable cloud managed by Amazon.com, which has suffered catastrophic data breaches in the recent past. In a survey, 40% of technology professionals said they had suffered a breach.

4. InBloom Inc. has already stated that it “cannot guarantee the security of the information stored... or that the information will not be intercepted when transmitted.”

(A-Z) Sitemap About this site Contact Legal Notice Search English
Privacy

• What is privacy?
  – Right to be let alone (Warren and Brandeis)
  – Informational self-determination (Westin)
  – Degree of access (Gavison)
  – ... Right to be forgotten ...

• Three dimensions (Roessler)
  – Informational privacy
  – Decisional privacy
  – Local privacy

• What it is not
  – Anonymity, secrecy, data protection
Prestaties werden ongevraagd verzameld en gegevens geanalyseerd.

De stichting Snippet, die tablets met onderwijs-apps verhuurt aan scholen, heeft de privacywetgeving geschonden, zegt het College Bescherming Persoonsgegevens (CBP).

Zonder dat scholen het weten verzamelde Snippet gegevens over de prestaties van de kinderen die met de tablets werkten. Die gegevens werden geanalyseerd en vergeleken met de prestaties van andere kinderen die het programma gebruikten. Ook werden ze gebruikt om de scholen ongevraagd advies te geven over de inhoud van hun lessen.
inBloom example(s) in the Netherlands

What are the dangers of learning analytics?

– Missing legal obligations:
  • Data protection
  • IRB
  • Education laws

– Inflicting harm:
  • Unfair discrimination
  • Unjustified discrimination (through errors)
  • Subjective privacy harm (panoptic effect)
  • Unintended pressure to perform / wrong incentives?
  • Anonymisation is not possible

– Violating human dignity

– Unintended changes of education norms?
Modernization of EU Universities report

Recommendation 14
Member States should ensure that legal frameworks allow higher education institutions to collect and analyse learning data. The full and informed consent of students must be a requirement and the data should only be used for educational purposes.

Recommendation 15
Online platforms should inform users about their privacy and data protection policy in a clear and understandable way. Individuals should always have the choice to anonymise their data.

#EP4LA on the European Agenda

- **Round table meeting ‘Ethiek en Learning Analytics’** (Jan 2014)
  https://www.surfspace.nl/media/bijlagen/artikel-1499-b315e61001041bf52a6b1c5d80053cea.pdf

- **Learning Analytics Summer Institute** (July 2014)
  http://lasiutrecht.wordpress.com/

- **Call for a ‘Code of Ethics for LA’** in NL (August 2014)
  https://www.surfspace.nl/artikel/1311-towards-a-uniform-code-of-ethics-and-practices-for-learning-analytics/

- **Call for a ‘Code of Ethics for LA’** in the UK (September 2014)
1st Call for Ethical & Privacy Issues in the Application of Learning Analytics

Organised by SURF SIG LA the https://www.surfspace.nl/sig/18-learning-analytics/ and the EU FP7 project LACE http://www.laceproject.eu Submission Deadline: 15.10.2014, Workshop will take place in Utrecht 28.10.2014, 11-5pm. The massive production, collection and processing of ...

Meer..

1. Privacy
2. Ethics
3. Data
4. Transparency

The rise of the #EP4LA project

- 1\textsuperscript{st} EP4LA @ Utrecht, NL, 28 October 2014
- 2\textsuperscript{nd} EP4LA @ Education Days, NL, 11 November 2014
- 3\textsuperscript{rd} EP4LA @ BDEDU, Washington, US, 11 November 2014
- 4\textsuperscript{th} EP4LA @ Apereo Foundation, FR, February 2015
- 5\textsuperscript{th} EP4LA @ JISC, February, UK, 2015
- 6\textsuperscript{th} EP4LA @ LAK15, NY, USA, March 2015
The Booksprint Method
What #EP4LA is aiming for

Do you want to input data into your own scientific/scholarly database, without sharing it with anyone other than your own team of researchers?

You do not need to seek consent for this.

Explanation: Even if the data is protected by copyright or database right, this type of use is one of the permitted exceptions: making a copy for your own use (Section 2.4.5) of a copyright-protected work or retrieving substantial portions of a database in the context of scientific/scholarly research (Section 3.2.5; retrieving non-substantial portions does not require consent in any case: Section 3.2.3).

Do you want to publish some of the data that you are borrowing or to share it with someone other than your own team of researchers?

Then you probably need to secure the author’s consent.

Check whether copyright applies:

Does the research data have an original character of its own and bear the personal stamp of the author? (Tip: Is it conceivable that two different authors, working separately, could have arrived at precisely the same form?.)

Copyright applies and you need to secure the consent of the author/authors.

Explanation: The decisive point is therefore whether the author had freedom of choice in arranging the data in a particular form and whether he did this – deliberately or otherwise – on the basis of personal preference (Section 2.3.2). Even just a selection consisting solely of bare facts may be subject to copyright protection (Section 2.3.3).
Example Submissions from Participants

- Who is in charge (who is the owners) of the data created by persons?
- What is the impact of privacy concerns for the management? How to deal with these concerns?
- Should students be allowed to opt-out of having their personal digital footprints harvested and analysed?
- How to prevent reuse of collected data for non-educational needs. (e.g. finance, insurance, research), or is it no problem?

We are practical people – our approach

• Invite 5 legal experts, rest Learning Analytics experts
• Task groups to answer requests of the stakeholders
Boundaries of Learning Analytics data

Where is the boundary on data use for learning analytics (courses, grades, LMS, GoogleDrive, library system, residence halls, dining halls, ...)?

– Contextual Integrity: context and norms of learning environment

– It depends on
  • Awareness of students about processes
  • Possible consequences for students
  • Safeguards that are in place
Outsourcing

What are the concerns when outsourcing the collection and analysis of data? Who owns the data?

– Concerns:
  • Undue third country data transfers
  • Less control about processing
  • Less transparency for the data subject

– Ownership:
  • No complete ownership for any party
  • Relevant: data protection and intellectual property rights
  • See discussion concerning ‘data portability’ in DP regulation (NDA agreement required)
Undesirable data collection

Are there any circumstances when collecting data about students is unacceptable/undesirable?

– Yes, there are:

• Data which is not of any purpose
• Data outside of the learning context
• Data of which the student is not aware
• Data which poses a risk to the student
• Data which is not well protected
Data access by students

What data should students be able to view, i.e. what and how much information should be provided to the student?

– Data Protection Directive (article 12):
  • Everything concerning them (at least upon request)
– Human subjects research:
  • Everything concerning study (at least after experiment)
  • Avoidance of deception
– But
  • Possible conflict of full data access with goals of LA?
  • How to provide meaningful access while excluding other students data?
9 Themes around privacy (1/3)

1. Legitimate grounds
- Why are you allowed to have the data?

2. Purpose of the data
- Repurposing is an issue vs. MIT Social Machine lab

3. Inventory of data
- What data do you have?
- What can you do with that data already?
9 Themes around privacy (2/3)

4. Data quality
- How good is the data? (eg. Bb log file is weak predictor)
- When do you I delete data and what data?

5. Transparency
- Informing students (Purpose, Approach)
- Checklist what to communicate for researchers

6. The rights of the data subject to access their data from the data client
- For teachers who are employees other rights apply
9 Themes around privacy (3/3)

7. Outsource processing to external parties
   - Prevent external parties to not do additional analysis (NDA agreement)

8. Transport data, legal location
   - e.g. Safe Harbour agreement

9. Data Security
   -> Shuangbao Wang
Solutions Privacy by Design

The 7 Foundational Principles
1. Proactive not Reactive;
2. Privacy as the Default Setting
3. Privacy Embedded into Design
4. Full Functionality
5. End-to-End Security
6. Visibility and Transparency
7. Respect for User

http://www.ipc.on.ca/images/resources/7foundationalprinciples.pdf
The Booksprint Rules

- All voices are valid
- Take ownership of the book (now and later)
- Session to be strongly facilitated

5 main elements:
1. Concept Mapping
2. Structuring
3. Writing
4. Composition
5. Publication

www.booksprints.net
A Twubs archive for this event is available at twubs.com/hashtag lak14.

Can help in report-writing and identifying those with similar interests.
De-anonymization of user data

Simple Demographics Often Identify People Uniquely

by Latanya Sweeney

In this document, I report on experiments I conducted using 1990 U.S. Census summary data to determine how many individuals within geographically situated populations had combinations of demographic value: combine in populations to uniquely or nearly uniquely identify individuals should not be considered anonymous. Yet, health surprising results using only three fields of information, even of 248 million) of the population in the United States had rep of birth}. About half of the U.S. population (132 million of 2 birth}, where place is basically the city, town, or municipality are likely to uniquely identify 18% of the U.S. population. In

L. Sweeney Simple Demographics Often Identify People Uni. (PDF).

movie ratings of 500,000 subscribers of Regina, the demonstrate that an adversary who knows only a little identify this subscriber’s record in the dataset. Us background knowledge, we successfully identified th apparent political preferences and other potentially so

![Figure 15 Linking to re-identify data](image-url)
Facebook Announces Stricter Guidelines For Research And Experiments On Its Users

Posted Oct 2, 2014 by Josh Constine (@joshconstine)
EDEN Conference

10 – 13 JUNE 2014

The Lanyrd social directory of events provides access to information about:

- Events
- Speaker
- Participants

The entry for this event may include access to:

- Slides
- Twitter archives
- Reports

Feel free to add your details

http://lanyrd.com/2014/eden14/
Practical considerations

- Data protection
- Anonymisation
- Human-subjects research
- (Contextual Integrity)
Data protection

• EU Data Protection Directive

• Common misconceptions
  – Scope of ‘personal data’
  – Relevance of consent
  – Anonymisation

• Legal ground: consent or legitimate interest?
• How to fulfil purpose limitation?
• Applicability of statistics exemption?
• Automated decisions?
Anonymisation

• Often considered as an ‘easy way out’ of DP obligations
• But
  – Other legal obligations?
  – Right to collect data in the first place?
  – Ethical considerations?
  – Unidentifiability achievable for dataset?

• Useful guidance: Opinion 05/2014 of Art. 29 WP
Human-subjects research

• Ethics approval required?
  – In NL: differs per university

• Likely problematic points:
  – Consent
  – Deception
  – Identifiable data
  – Subordinate position of students
Transparency

• Transparency about
  – Data collection
  – Profiling activities
  – Use of results within institutional process

• Data Protection directive
• Human-subjects research

• Failure of notice and consent
• Legal texts vs. layered notices vs. icons vs. data cockpits
Value Sensitive Design (Batya Friedman)

- Goal: address human values in a technical design

Source: presentation by Jeroen van den Hoven
Reference that are relevant
Formal frameworks

• Contextual Integrity

• Value Sensitive Design
Contextual Integrity (Helen Nissenbaum)

• Benchmark for privacy in information systems
  - Recognition of social contexts
  - Related to ‘reasonable expectation’

• Contextual integrity ≈ fulfilling informational norms
  - Contexts: roles, activities, norms, values
  - Actors: senders, receivers, subjects
  - Attributes: data fields
  - Transmission principles: confidentiality, etc.
Data Protection aka Directive 95/46/EC

- Privacy as control / informational self-determination
- Not necessarily restricting data collection, but channelling it

- In the European Union
  - Directive 95/46/EC

- Elsewhere
  - Fair Information Practices (US, not mandatory)

- Upcoming: general data protection regulation
95/46/EC – general obligations
(article 6 (1))

a) Fairly and lawfully
b) Purpose limitation (see also Opinion 03/2013 of Art 29 WP)
   • Specified, explicit and legitimate
   • Compatible use (no secondary use)
   • Exceptions for statistics and science
c) Adequate, relevant and not excessive
d) Accurate and up to date
e) Identification no longer than necessary

• Important: these provisions also hold if consent of data subject has been collected!
95/46/EC – legal grounds (article 7)

a) Consent
b) Performance of a contract
c) Legal obligations
d) Vital interests of the data subject
e) Public interest
f) Legitimate interests

- Legitimate interest should be used more often
- ‘how to’: see Opinion 04/2014 of Art 29 WP
  - Balancing test: nature of the interest, impact on data subject, additional safeguards
95/46/EC - further considerations

• Special categories of data require extra care (article 8)
  – ethnic origin, political opinion, etc.
• Right of access to data (article 12)
  – Includes rectification
• Automated decisions (article 15)
  – Decisions severely affecting the data subject may not be solely based on automated profiling
  – If so, appeal to decision must be possible
• Security obligation (article 17)
• Notification obligation (article 18)
• Third country data transfers (article 25)
  – Only allowed if adequate level of protection
  – For the US: Safe Harbour arrangement
95/46/EC - scope

• Only applicable if
  – automated processing of ‘personal data’ OR
  – part of a ‘filing system’ (article 3)

• Personal data
  – Natural person is directly or indirectly identifiable

• → properly anonymised data not subject to 95/46/EC

• However,
  – Other legal obligations exist
    • e.g. article 7 and 8 EU Charter of Fundamental Rights, article 8 ECHR, e-Privacy Directive
  – Anonymous data can still inflict harm (→ group profiling)
  – Anonymisation itself is also data processing!
Anonymisation – legal considerations

1/2

• Good introduction in Opinion 05/2014 of Art 29 WP

• *identification no longer than necessary* can be an obligation for anonymisation
  • → anonymisation by default?

• The process of anonymisation is also data processing
  • Principle of purpose limitation applies!

• Applying anonymised data to user profiles is data processing
  • → only the processing of already anonymised data by itself is not subject to data protection
• Pseudonymisation is not anonymisation
  – replacing identifiers is not enough

• Only anonymous if identification is irreversibly prevented
  – Whether that is the case depends on robustness of method
  – Opinion 05/2014 assesses different methods in that regard

• Residual risk of identification has to be taken into account
  – Monitor and control
Anonymisation – Practical considerations

• Decisive is the ability to single out an individual
  – Dropping identifier columns from a DB is likely not enough
  – Attributes such as age, gender, etc. might have to be modified
  – →randomization and generalization

• Consider the possibility of further harm (group profiling)
  – Unjustified discrimination
  – Unfair discrimination
Human-subjects research 1/2

- Experiments with Learning Analytics might require approval of a local ethics committee!

- Nuremberg code, Declaration of Helsinki (excerpt)
  - Informed consent
  - For the good of society
  - Avoidance of harm
  - Option to exit

- Problematic points
  - Consent
  - Deception
  - Identifiable data
  - Subordinate position
Human-subjects research 2/2

• Required by law only for medical research
• Regulations differ per university

• If review is necessary, often:
  – Multiple stages
  – Checklists
  – Possibly advice for improvement of study

• Requirements for approval tie in with ethical considerations discussed earlier
Conclusions

• No simple recipe or ‘how to ethics’

• Data protection ≠ anonymisation or consent
  • Protecting privacy ≠ anonymisation

• Ethics in LA is more than privacy
  • Legal constraints (data protection) are not everything

• Think along the lines of contextual integrity
  • Keep the students involved
Discussion

• Limits of privacy by design
• Security best practices
• Difficulties of meten is weten
  – Discussion around publication of CITO scores
• Dangers of misinterpretation
  – Safeassign Matching Score
• Data access
  – Student, teacher or both
  – Granularity of access
• Connection to non-university systems
  – Cloud services, social media
• Ways forward
  – Code of ethics for LA?
“Ethics & Privacy Issues in the Application of Learning Analytics” by Hendrik Drachsler, Open University of the Netherlands was presented at LACE & SURF workshop, Utrecht, Netherlands on 28.10.2014.

Hendrik.drachsler@ou.nl, @hdrachsler

This work was undertaken as part of the LACE Project, supported by the European Commission Seventh Framework Programme, grant 619424.

These slides are provided under the Creative Commons Attribution Licence: http://creativecommons.org/licenses/by/4.0/. Some images used may have different licence terms.
You are free to:

- copy, share, adapt, or re-mix;
- photograph, film, or broadcast;
- blog, live-blog, or post video of

this presentation provided that:

You attribute the work to its author and respect the rights and licences associated with its components.