WEBS6201 Foundations of Web Science

Week 2
The Data Pipeline
Mark Weal
Digital Data Infrastructures: interrogating the social media data pipeline.

Les Carr and Susan Halford, Web Science Institute, University of Southampton.

Based on a paper by Susan Halford, Mark Weal, Ramine Tinati, Les Carr and Catherine Pope in New Media and Society, http://journals.sagepub.com/doi/abs/10.1177/1461444817748953
Social Media Data

• From social media practices and effects ... to data
• An unexpected gift bringing rich research opportunities
• Enthusiasm - the telescope - the macroscope ...
  ‘... it is as if the inner workings of private worlds have been pried open’ (Latour 2007)
• Scepticism:
  ‘[w]hatever value big data may have for “knowing capitalism”, its’ value to social science has ... [f]or the present at least, to remain very much open to question’ (Goldthorpe 2016)
Theorising data

• No such thing as ‘raw’ or ‘naturally occurring’ data
• All data are ‘always already social’ (Bowker 2013)
• We must explore:
  ‘… the lives and specificities of devices and data themselves, where and how
  they happen, who and what they are attached to and the relations they forge, how
  they get assembled, where they travel, their multiple arrangements and
  mobilizations and, of course their instabilities, durabilities and how they
  sometimes get disaggregated too’ (Ruppert, Law & Savage 2013)
• Where to start?
A methodology

We have conducted a survey of 10,000 people. We don't know how these 10,000 people were selected as the company that collected the data for us won't share this information, but we know that at some point they said something about Copenhagen or Aarhus. These may not be 10,000 unique individuals; we have no way of knowing. Some people may have filled the form in multiple times. Some of the forms (up to 20%) may have been filled in by computer algorithms. We don't know which is which. Different people have filled in different variations of the form. We don't know which is which.

We have categorised people into those from Copenhagen and those from Aarhus based on location information they have either supplied (very rarely), or that we have inferred.

We have analysed the data using a small, not exhaustive, quite subjective set of positive and negative words to give each person a happiness rating from 1-10. Based on this we can say with a $p=0.05$ level of certainty that people living in Copenhagen are happier than people living in Aarhus.
Another methodology

“For this paper, a number of interviews had been conducted, transcribed and printed by other researcher(s) for unknown reasons. The resulting printouts were torn into shreds, mixed up and dumped. I have used some of these found statements and interpreted them against my research questions.”
A middle path, between giving in and getting out (Gehl 2015)
The Data Pipeline

- Sociotechnical
- Iterative
- Core to the generation of data
- Core to the circulation of data
- Methodological implications?
1: Population

• Demographics
• Location

GPS location enabled - <3%
Jakarta 2.86%
Moscow 0.77%
1: Population

- Demographics
- Location
- Users – sovereign individuals?
• How the data are harvested shapes the sample
• The API shapes the sample e.g. % data streams, real time/historic
• Rate limiting
3: Method

- Instruments for data collection
- Affordances
- For example: functionalities, data bases – shape data in specific ways over time
<table>
<thead>
<tr>
<th>Database</th>
<th>Storage design and method shapes the types of information recorded about users.</th>
<th>Historic data storage decisions and technical query limitations may shape what data are included in samples.</th>
<th>Considerations of cost, performance and business requirements for data storage may shape what data are collected and stored and how.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server Software</td>
<td>Determines who or what has access to the service, and what information is required to set up an account.</td>
<td>Server capacity may restrict data volume delivered; geographical location of server may affect data delivered.</td>
<td>Operates data management (e.g. spam removal and moderation, load balancing) shaping what data are collected.</td>
</tr>
<tr>
<td>API</td>
<td>APIs may not recognise all characters (languages) effectively; or be available to all operating systems/software development toolkits</td>
<td>A variety of differently structured samples may be available.</td>
<td>Defines the scope and volume of what data can be collected, stored and queried.</td>
</tr>
<tr>
<td>Harvesting Method</td>
<td>Harvesting methods construct different views of the populations. Web scraping may be more likely to access the population of currently active users, which could be different to the population accessed via historical searches using an API.</td>
<td>Web scraping will by-pass ‘official’ data samples, offering data from a sample of web pages. This sample may be affected by the ‘filter bubble’ of the person accessing the web pages. Use of third party data may introduce additional sampling effects.</td>
<td>Different harvesting methods have access to different types of data about the population and sample.</td>
</tr>
<tr>
<td>Client Software</td>
<td>Different clients may generate different information about the population. On some platforms you may know what client generated the content (this used to be the case on Twitter), on many though you can’t know this.</td>
<td>Some clients (apps) may receive more data than others (if harvesting through a client).</td>
<td>Different clients may produce distinctive forms of data and metadata e.g. some may add geographic data by default, some might link directly to shared or re-shared material.</td>
</tr>
<tr>
<td>Subject</td>
<td>Different subjects – human/non-human, demographically distinct – may characterise particular platform populations.</td>
<td>User activities may shape sampling methods (e.g. official samples may focus on central or highly active users.)</td>
<td>User practices and meanings shape the data generated and the claims that can be made from these.</td>
</tr>
</tbody>
</table>
Framework of social media engagement (1-4)

• Presentation of self, reputation management, and privacy
  – Identity is crafted through the development of a personal profile or virtual self over time on social media.

• Action and participation
  – Social media enables users to perform a variety of activities such as viewing, posting, or sharing content, collaboration, and discussion.

• Uses and gratifications
  – Social media users have different motivations for adoption and use including, for example, the exchange of information and the social benefits derived from its use.

• Positive and negative experiences
  – positive aspects of social media compel people to use it such as positive emotions, serendipity, and flow.
  – negative experiences such as spam, fraud, cyberbullying, misogyny, trolling may lower user engagement with social media.


Framework of social media engagement (5-7)

- Usage and activity counts
  - Usage and activity counts refer to the data associated with users’ actions and participation within a particular social media site [aka analytics], which may be presented in real-time in raw or aggregate form to users.

- Social context
  - the social, political, economic, work, and personal phenomena or characteristics that underlie a users’ social networks within social media sites, including the size and nature of these local and global networks (e.g. a small, close-knit peer group; a large, diffuse network of social activists)

- Platform characteristics
  - Features of specific platforms may influence engagement – e.g. features that enable users to share information or communicate directly with one another.


Police.uk

http://www.police.uk
Police.uk

• What are they trying to achieve by putting this information on a map?
Privacy and Anonymisation

Trying to find a balance between providing granular crime data and protecting the privacy of victims has been one of the biggest challenges involved in releasing this data.

We consulted heavily with the Information Commissioner's Office and Data Protection specialists in the Home Office in the run up to releasing this data, and worked within their guidance to create an anonymisation process which adequately minimises privacy risks whilst still meeting our transparency goals and being useful to the public.

We continue to work with the ICO and data users to improve the anonymisation processes and make sure that we have the balance right. You can see some of the improvements that we've made over the previous years on the changelog page.

http://data.police.uk/about/#anonymisation
## Anonymisation methods

<table>
<thead>
<tr>
<th>Field</th>
<th>Raw Example</th>
<th>Anonymisation Method</th>
<th>Anonymised Example</th>
<th>Published in CSVs</th>
<th>Published in API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offence Reference</td>
<td>C2/0123/13</td>
<td>Replaced with a unique one-way hash of the record.</td>
<td>7512cb691a6f9c6c47a2c dfdcd0a1f8735a0870d3a 2bde21d8311bd74f17eeeb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>2013-07-23</td>
<td>Truncated to show the year and month only.</td>
<td>2013-07</td>
<td>Month</td>
<td>date</td>
</tr>
<tr>
<td>Home Office Offence Code</td>
<td>104/25</td>
<td>Assigned into one of 15 categories. A complete mapping between Home Office Offence Codes and Categories can be downloaded here.</td>
<td>Violence and Sexual Offences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easting</td>
<td>519500</td>
<td>Anonymised in line with the location anonymisation process. Converted to WGS84 latitude.</td>
<td>52.58019</td>
<td>Latitude</td>
<td>longitude</td>
</tr>
<tr>
<td>Northing</td>
<td>299500</td>
<td>Anonymised in line with the location anonymisation process. Converted to WGS84 longitude.</td>
<td>-0.23782</td>
<td>Longitude</td>
<td>longitude</td>
</tr>
</tbody>
</table>
Crime types (15 categories)

- Violence and Sexual Offences
- Criminal Damage and Arson
- Drugs
- Burglary
- Robbery
- Public Order
- Vehicle Crime
- Shoplifting
- Bicycle Theft
- Other Theft
- Other Crime
Other Crimes

- Forgery of a drug prescription or copying a false drug prescription
- Possess false mental health document
- Supplying false information to assay office
- Keeping a brothel used for prostitution
- Trafficking people into the UK for the purpose of exploitation
- Failure to withdraw from circulation a Euro note or coin which is believed to be counterfeit
- Supplying article with unauthorised mark
- Assisting a prisoner to escape
- Melting down or breaking up metal coin without licence
- False written statements tendered in evidence
- Receiving stolen goods
- Concealment of birth
- Going equipped for stealing etc.
- Causing or inciting prostitution for gain
- Bigamy
- ...
Location anonymisation

How are crime locations anonymised?

We maintain a master list of over 750,000 'anonymous' map points. Each map point is specifically chosen so that it:

- Appears over the centre point of a street, above a public place such as a Park or Airport, or above a commercial premise like a Shopping Centre or Nightclub.
- Has a catchment area which contains at least eight postal addresses or no postal addresses at all.

When crime data is uploaded by police forces, the exact location of each crime is compared against this master list to find the nearest map point. The co-ordinates of the actual crime are then replaced with the co-ordinates of the map point. If the nearest map point is more than 20km away, the co-ordinates are zeroed out. No other filtering or rules are applied.
Conclusion

• Recognise the limits of what we can and can’t know about social media data

• Key steps

(1) Transparency
(2) Consider implications of data construction for research questions
(3) Knowledge claims
(4) Creative data assemblages