DCS/CSCI 2350: Social & Economic Networks

Privacy & Influence in Online Social Networks: Case Study of Facebook & Cambridge Analytica

Mohammad T. Irfan
Events (and behind the scene)

What happened (2014 – 2016)

Aleksandr Kogan
Professor at Cambridge U
Timeline

2010 - Launches Open Graph API for app developers
2011 - Agreement with FTC: Consent for sharing user data
2013 - CA founded in the UK to operate in the US
2013 - Makes a Facebook personality quiz app for academic research
2014 - Forms a company (GSR) to harvest Facebook data
2014 - Sells data to Cambridge Analytics
2014–16 - Works for Brexit and Trump

How it came to light

- Christopher Wylie: whistleblower
- March 17 & 18: sensational reporting by Carole Cadwalladr on NYTimes and the Guardian
Behind the scene

As told by Carole Cadwalladr


Prof. Kogan & Cambridge University

Prof. Kogan: blamed by Facebook for violating ToS

University colleagues upset

Prof John Rust, the director at Cambridge University’s Psychometrics Centre, ... accused Kogan of trying to make $1m in “personal profit in terms of asset and data” from the scheme, while only reimbursing his fellow psychologists, Dr Michal Kosinski and Dr David Stillwell, who had led much of the cutting-edge research, with $100,000.


Prof. Kogan’s defense

Behind the scene: Cambridge Analytica (CA)

- Alleged business practices (denied by CA)
  - [https://www.youtube.com/watch?v=mpbeOCKZffQ](https://www.youtube.com/watch?v=mpbeOCKZffQ)

Behind the scene: Facebook

- August 2016: Facebook lawyers asked all parties to delete the data
- Everyone certified deletion
- But did they actually delete it?
  - March 2018: News channel obtains data of 136,000 Colorado residents
- Facebook kept denying data/trust breach till March 2018
Data and Computation

What does the data look like?

- Courtesy of Prof. David Carroll (Parsons New School)
- Voter data
  - “accurate and mostly complete for me”

Partial snapshot

<table>
<thead>
<tr>
<th>primary_residence_US_Congressional_District</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary_residence_State_Senate_District</td>
</tr>
<tr>
<td>primary_residence_State_House_District</td>
</tr>
<tr>
<td>primary_residence_State_Legislative_District</td>
</tr>
<tr>
<td>primary_residence_Designated_Market_Area_DMA</td>
</tr>
<tr>
<td>primary_residence_County</td>
</tr>
<tr>
<td>primary_residence_Voters_FIPS</td>
</tr>
<tr>
<td>primary_residence_Borough</td>
</tr>
<tr>
<td>primary_residence_Village</td>
</tr>
<tr>
<td>primary_residence_Township</td>
</tr>
<tr>
<td>primary_residence_Town_District</td>
</tr>
<tr>
<td>primary_residence_Precinct</td>
</tr>
<tr>
<td>primary_mailing_AddressLine</td>
</tr>
<tr>
<td>primary_mailing_ExtraAddressLine</td>
</tr>
<tr>
<td>primary_mailing_City</td>
</tr>
<tr>
<td>primary_mailing_State</td>
</tr>
<tr>
<td>primary_mailing_Zip</td>
</tr>
</tbody>
</table>
What does the data look like?

- Predictive scores for Prof. Carroll
- “Feels roughly accurate.”

<table>
<thead>
<tr>
<th>Issue Category</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Debt Importance Rank 1-10</td>
<td>10</td>
</tr>
<tr>
<td>Gun Rights Importance Rank 1-10</td>
<td>3</td>
</tr>
<tr>
<td>Traditional Social and Moral Values Importance Rank 1-10</td>
<td>9</td>
</tr>
<tr>
<td>Environment Importance Rank 1-10</td>
<td>5</td>
</tr>
<tr>
<td>Education Importance Rank 1-10</td>
<td>4</td>
</tr>
<tr>
<td>National Security Importance Rank 1-10</td>
<td>7</td>
</tr>
<tr>
<td>Immigration Importance Rank 1-10</td>
<td>8</td>
</tr>
<tr>
<td>Socially Progressive Civil Rights Importance Rank 1-10</td>
<td>6</td>
</tr>
<tr>
<td>Jobs and Economy Importance Rank 1-10</td>
<td>1</td>
</tr>
<tr>
<td>Healthcare Importance Rank 1-10</td>
<td>2</td>
</tr>
<tr>
<td>Registered Partisanship</td>
<td>Very Unlikely Republican</td>
</tr>
<tr>
<td>Unregistered Partisanship</td>
<td>Very Unlikely Republican</td>
</tr>
<tr>
<td>2016 General Election Turnout Propensity</td>
<td>Very High</td>
</tr>
</tbody>
</table>

But... is this his real data?

Carole Cadwalladr 🈴️
@carolecadwalla

Here's what it knows about @profcarroll
He asked for his data back using UK laws
twitter.com/profcarroll/st ...
Only this isn't his "real" data...

David Carroll 👤 @profcarroll
Replying to @profcarroll

7/ Here is CA/SCL voter issue profile/propensity model on me. I'd rank this somewhat differently but feels roughly accurate. Could be worse.
What algorithms are used?

- Leaked codes from Aggregate IQ
  - [https://www.upguard.com/breaches/aggregate-iq-part-one](https://www.upguard.com/breaches/aggregate-iq-part-one)
- Data collection
- Algorithm to go from quiz-takers to their friends, friends-of-friends, etc.
  - 75 million nodes starting with just 250,000

From data to prediction

- Computational problem: Given Facebook data, predict the OCEAN scores
  - Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism
- Machine learning
Machine learning: big picture

- Training data: Facebook activities and "true" OCEAN scores
- Machine learning algorithm
- Test data: Facebook data of a new person
- Model of classification/regression
- Prediction: OCEAN scores

Issues and Implications
Issues
- Privacy in digital life
- Psychographic modeling
- Computational issues: local → global
- Microtargeting and influence in online world
- Political campaigns
- Ethics in research
- Legal issues
- Public relations

Writing assignment
- The Tipping Point
  - Give mathematical models
  - Extend existing models
- Is Gladwell's thesis valid in the online world?
  - Three laws and how/whether they apply to our digital life
    - Law of the few: connectors, mavens, salesmen
    - Stickiness factor
    - Power of context