Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski et al.

The Psychometrics Centre, University of Cambridge, UK

Computational Social Media EPFL Course
Rémi Lebret
7th March 2014
Introduction
Digital records of individuals behavior

- People may choose not to reveal certain pieces of information about their lives.
  - sexual orientation, age, ...
- This information might be predicted.
  - Predict pregnancies from customer shopping records.
- It could lead to tragic outcome.
  - Revealing pregnancy of an unmarried woman to her family in a culture where this is unacceptable.
  ⇒ Dangerous invasion of privacy.
Predicting individual traits and attributes

From:

- People’s Web site browsing logs
- Contents of personal Web sites
- Music collections
- Properties of Facebook or Twitter profiles
- Language used by users
- Location within a friendship network at Facebook → Predictive of sexual orientation
Facebook Likes

Mechanism used by Facebook users to express their positive association with online content.

- Very generic class of digital records.
  → Similar to Web search queries, Web browsing history, and credit card purchases.
- Currently public available by default.

→ Automatically and accurately estimate personal attributes using these basic records.
Method
Design of the study

Selection of traits and attributes potentially private:

- Sexual orientation
- Ethnic origin
- Political views
- Religion
- Personality
- Intelligence
- Satisfaction with life (SWL)
- Substance use (alcohol, drugs, cigarettes)
- User’s parents together at 21
- Basic demographic attributes
Design of the study

Selection of traits and attributes potentially private:

- Sexual orientation
- Ethnic origin
- Political views
- Religion
- Personality → International Personality Item Pool questionnaire
- Intelligence
- Satisfaction with life (SWL)
- Substance use (alcohol, drugs, cigarettes)
- User’s parents together at 21
- Basic demographic attributes
Design of the study

Selection of traits and attributes **potentially private**:

- Sexual orientation
- Ethnic origin
- Political views
- Religion
- Personality → International Personality Item Pool questionnaire
- Intelligence → Raven’s Standard Progressive Matrices (SPM)
- Satisfaction with life (SWL)
- Substance use (alcohol, drugs, cigarettes)
- User’s parents together at 21
- Basic demographic attributes
Design of the study

Selection of traits and attributes potentially private:

- Sexual orientation
- Ethnic origin
- Political views
- Religion
- Personality → International Personality Item Pool questionnaire
- Intelligence → Raven’s Standard Progressive Matrices (SPM)
- Satisfaction with life (SWL) → SWL scale (5-item instrument)
- Substance use (alcohol, drugs, cigarettes)
- User’s parents together at 21
- Basic demographic attributes
Selection of traits and attributes potentially private:

- Sexual orientation
- Ethnic origin
- Political views → Facebook profiles
- Religion → Facebook profiles
- Personality → International Personality Item Pool questionnaire
- Intelligence → Raven’s Standard Progressive Matrices (SPM)
- Satisfaction with life (SWL) → SWL scale (5-item instrument)
- Substance use (alcohol, drugs, cigarettes)
- User’s parents together at 21
- Basic demographic attributes → Facebook profiles
Design of the study

Selection of traits and attributes **potentially private:**

- Sexual orientation
- Ethnic origin
- Political views → Facebook profiles
- Religion → Facebook profiles
- Personality → International Personality Item Pool questionnaire
- Intelligence → Raven’s Standard Progressive Matrices (SPM)
- Satisfaction with life (SWL) → SWL scale (5-item instrument)
- Substance use (alcohol, drugs, cigarettes) → Online survey
- User’s parents together at 21 → Online survey
- Basic demographic attributes → Facebook profiles
Design of the study

Selection of traits and attributes potentially private:

- Sexual orientation
- Ethnic origin → Visual inspection of profile pictures
- Political views → Facebook profiles
- Religion → Facebook profiles
- Personality → International Personality Item Pool questionnaire
- Intelligence → Raven’s Standard Progressive Matrices (SPM)
- Satisfaction with life (SWL) → SWL scale (5-item instrument)
- Substance use (alcohol, drugs, cigarettes) → Online survey
- User’s parents together at 21 → Online survey
- Basic demographic attributes → Facebook profiles
Design of the study

Selection of traits and attributes potentially private:

- Sexual orientation → Facebook profile “Interested in” field
- Ethnic origin → Visual inspection of profile pictures
- Political views → Facebook profiles
- Religion → Facebook profiles
- Personality → International Personality Item Pool questionnaire
- Intelligence → Raven’s Standard Progressive Matrices (SPM)
- Satisfaction with life (SWL) → SWL scale (5-item instrument)
- Substance use (alcohol, drugs, cigarettes) → Online survey
- User’s parents together at 21 → Online survey
- Basic demographic attributes → Facebook profiles
Sample

- 58,466 US Facebook users
- Psychodemographic profile + list of their Likes
- From myPersonality Facebook application
Prediction with Facebook Likes

1. Users’ Facebook Likes
   - 55,814 Likes
   - User – Like Matrix (10M User-Like pairs)

2. Singular Value Decomposition
   - 100 Components
   - User – Components Matrix

3. Prediction Model
   - Using Logistic or Linear Regression (with 10-fold cross validation)
   - e.g. \( \text{age} = \alpha + \beta_1 C_1 + \ldots + \beta_n C_{100} \)
   - Predicted variables
     - Facebook profile: age, gender, political and religious views, relationship status, proxy for sexual orientation, social network size and density
     - Profile picture: ethnicity
     - Survey/test results: BIG5 Personality, intelligence, satisfaction with life, substance use, parents together?
Results
**Prediction accuracy**

(a) Accuracy of classification for dichotomous attributes.

(b) Accuracy of regression for numeric attributes and traits.
Amount of Data Available and Prediction Accuracy

- Between 1 and 700 Likes per individual → median = 68
## Selected Most Predictive Likes

<table>
<thead>
<tr>
<th>IQ</th>
<th><strong>High</strong></th>
<th><strong>Low</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Godfather</td>
<td>Sephora</td>
</tr>
<tr>
<td></td>
<td>Mozart</td>
<td>I Love Being A Mom</td>
</tr>
<tr>
<td></td>
<td>Lord Of The Rings</td>
<td>Harley Davidson</td>
</tr>
<tr>
<td>Extraversion</td>
<td><strong>Outgoing &amp; Active</strong></td>
<td><strong>Shy &amp; Reserved</strong></td>
</tr>
<tr>
<td></td>
<td>Beerpong</td>
<td>Video Games</td>
</tr>
<tr>
<td></td>
<td>Dancing</td>
<td>Programming</td>
</tr>
<tr>
<td></td>
<td>Cheerleading</td>
<td>Role Playing Games</td>
</tr>
<tr>
<td></td>
<td>Chris Tuker</td>
<td>Manga</td>
</tr>
<tr>
<td></td>
<td>Theatre</td>
<td>Minecraft</td>
</tr>
<tr>
<td>Openness</td>
<td><strong>Liberal &amp; Artistic</strong></td>
<td><strong>Conservative</strong></td>
</tr>
<tr>
<td></td>
<td>Oscar Wilde</td>
<td>NASCAR</td>
</tr>
<tr>
<td></td>
<td>Leonardo Da Vinci</td>
<td>ESPN2</td>
</tr>
<tr>
<td></td>
<td>Leonard Cohen</td>
<td>The Bachelor</td>
</tr>
<tr>
<td></td>
<td>Bauhaus</td>
<td>Justin Moore</td>
</tr>
</tbody>
</table>
Average Levels for several popular Likes
Average Levels for several popular Likes

Democrat
- Barack Obama
- The Colbert Report
- Nicki Minaj
- Cuddling
- Hello Kitty

Homosexual
- The Colbert Report
- Barack Obama
- Cuddling

Christian
- Mitt Romney
- Nicki Minaj
- Hello Kitty
- Cuddling
- Camping

African-American
- Nicki Minaj
- Hello Kitty
- Barack Obama
- Cuddling
- Camping
Conclusion

Variety of people’s personal attributes can be inferred using their Facebook Likes.

+ Improve numerous products and services
  - Permit assessment across time to detect trends
  - Open new doors for research in human psychology

- No individual consent, no notice
  - Infer attributes that individual intends not to share
  - Threat to individual’s well-being, freedom or even life

Trust in online services could decrease without transparency and control over users informations.