YOU ARE WHAT YOU LIKE
INFORMATION LEAKAGE THROUGH
USERS’ INTERESTS

Abdelberi (Beri) Chaabane, Gergely Acs, Mohamed Ali Kaafar
Most visited websites:
- Facebook (2nd), YouTube (3rd), Twitter (10th)

Facebook¹:
- > 800M users
- > 350M users access through their mobile
- > 250M photos are uploaded per day
- > 20M application installation per day

And privacy ??

Identifying the threat

Users' private / pub data

Mark Z. is a bad guy!

Privacy Policies

~ Private Profiles

Inference Technique

Identifying the threat
Goal

- Inferring Missing/Hidden information from a public user profile
  - Using Friendship or links information\textsuperscript{[2,3]}
  - Only using user’s revealed data

*: http://13thfloorgrowingold.wordpress.com/*
What people reveals?

<table>
<thead>
<tr>
<th>Topic</th>
<th>25%</th>
<th>75%</th>
<th>21%</th>
<th>79%</th>
<th>43%</th>
<th>57%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Current City Looking for

<table>
<thead>
<tr>
<th>Topic</th>
<th>23%</th>
<th>77%</th>
<th>22%</th>
<th>78%</th>
<th>22%</th>
<th>78%</th>
<th>17%</th>
<th>83%</th>
<th>16%</th>
<th>84%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthday</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hometown

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Interested In</th>
</tr>
</thead>
<tbody>
<tr>
<td>78%</td>
<td>84%</td>
</tr>
<tr>
<td>78%</td>
<td>84%</td>
</tr>
<tr>
<td>78%</td>
<td>84%</td>
</tr>
<tr>
<td>83%</td>
<td>84%</td>
</tr>
</tbody>
</table>

Missing values

<table>
<thead>
<tr>
<th>22%</th>
<th>78%</th>
<th>22%</th>
<th>78%</th>
<th>17%</th>
<th>83%</th>
<th>6%</th>
<th>1%</th>
<th>94%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>77%</td>
<td>78%</td>
<td>78%</td>
<td>78%</td>
<td>83%</td>
<td>99%</td>
<td>94%</td>
<td>1%</td>
<td>6%</td>
<td>1%</td>
</tr>
</tbody>
</table>
Quiz

Who is this guy? Who likes his music?
Music? Why would that work?

- In real life, an individual interest (or lifestyle) might reveal many aspects of his personal information
  - demographics or geopolitical aspects.

- Availability
  - Seemingly harmless ;-) 
  - by default settings?
Not that easy

- Heterogeneity
  - Too general “I like Jazz Music”
  - Too specific “Angus Young”

- Difficult to semantically link interests
  - What is the link between Angus Young, Brian Johnson and High Voltage?
One of the MOST available data

Describe users’ tastes

Can be used to derive user information
  - Gender, Location, Age, Marital status, Religion, etc.

Very sparse (millions of likes)

User-generated (No defined pattern)

No “standard” granularity
A toy example

Mohammad-Reza Shajarian, Nazeri, Gogosh

What does it mean (lack of semantics)

What can we infer?
Semantics: a naïve example

- **Shajarian**: 1940 births; Living people; Iranian classical; vocalists Iranian; humanitarians Iranian; male singers; Iranian musicians

- **Nazei**: Grammy Award winners; Iranian Kurdish people; Living people; Iranian classical vocalists; Iranian humanitarians; Iranian Légion d’honneur recipients; Iranian male singers

- **Gogosh**: people of Azerbaijani descent; Iranian female; Persian-language singers; Iranian pop singers; Iranian Shi’a; Muslims People from Tehran

Btw it belongs to
http://facebook.com/kave.salamatian
Semantics: a naïve example II

- **Shajarian**: 1940 births; Living people; **Iranian** classical; vocalists **Iranian**; humanitarians **Iranian**; male singers; **Iranian** musicians

- **Nazei**: Grammy Award winners; **Iranian Kurdish** people; Living people; **Iranian** classical vocalists; **Iranian** humanitarians; **Iranian** Légion d'honneur recipients; **Iranian** male singers

- **Gogosh**: people of **Azerbaijani**; descent **Iranian** female; **Persian-language** singers; **Iranian** pop singers; **Iranian** Shi'a; **Muslims** People from Tehran
The Algorithm

Step 1: Extract Semantics

Step 2: Compute Interest Feature Vectors (IFV)

Step 3: Classify Users
Infer Attributes
Step 1
Tree of wikipedia

- Fundamental
  - Concepts
    - children
  - Life
    - children
  - Matter
    - children
  - Society
    - Communication
      - Mass Media
      - Social networks
        - Social Network services
        - Facebook
        - ...

Extract semantic (Description)

- ‘Ontologized’ version of wikipedia
  - Using the “structured knowledge” of Wikipedia
    - Extract keywords from a certain ‘granularity’

- Each like is an article

- Extract Parent Categories of the ‘like’ article
  - Using the same granularity
Using the same granularity allows us to semantically ‘link’ similar concepts.

**AC/DC:** Australian heavy metal musical groups; Australian hard rock musical groups; Blues rock groups; Musical groups established in 1973;

**Angus Young:** AC/DC members; Australian blues guitarists; Australian rock guitarists; Australian heavy metal guitarists

**High Voltage:** AC/DC songs; Songs written by Angus Young; 1970s rock song stubs
The Algorithm

Step 1: Extract Semantics

Step 2: Compute Interest Feature Vectors (IFV)

Step 3: Classify Users
Infer Attributes
Step 2
LDA Intuition

**All available Interests**

Interest1: \( w_1, w_2, w_3 \ldots \)

**LD.**

Classify

**Topic 1:**

\[ \text{Prob}(I_1 \rightarrow T_1) \]

\[ \text{Prob}(I_2 \rightarrow T_1) \ldots \]

**K topics**

\( I_1: \) Interest1

\( T_1: \) Topic 1
LDA as a Probabilistic model

1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.

2. Infer the hidden structure using posterior inference
   - What are the topics that describe this collection?

3. Situate new data into the estimated model.
   - How does this new document fit into the estimated topic structure?

D.Blei (MLSS’09)
Words collected into documents
- Each document is a mixture of a small number of topics
- Each word's creation is attributable to one of the document's topics
- Topics are not nominative

Input:
- Documents (words Frequency)
- Number of Topics (K)

Output
- Word distribution per topic
- Probability for each document to belong to each topic
Topic example
The Algorithm

Step 2: Compute Interest Feature Vectors (IFV)

Step 3: Classify Users
Infer Attributes
Step 3
Inferring Hidden Attribute

- IFV ‘uniquely’ quantifies the interest of each user along topics

- Classify users based on IFV
  - Simple approach
  - Using the nearest neighbors (K-NN)

- Similar users grouped together.
  - User sharing the ‘same’ taste should share the same attributes
Nearest Friend Neighbor

- We define an appropriate distance measure in this space: chi-squared distance metric

\[ d_{V,W} = \sum_{i=1}^{k} \frac{(V_i - W_i)^2}{(V_i + W_i)} \]

- Using Kd-tree to reduce the computation from \( M^2 / 2 \) to \( O(M \log_2 M) \)
The \( n \) nearest users to user1 are: \( S=\{\text{user3, userm, } \ldots\} \)

The attribute is equal the the **majority** of the attribute in \( S \) (Majority voting)
Datasets

- **Public Profiles**
  - Crawled more than 400k profiles (Raw-Profiles)
  - More than 100k Latin-written profiles with music interests (Pub-Profiles)

- **Private Profiles**
  - Using a Facebook App.
  - More than 4000 Private profiles (used 2.5 K, Volunteer-Profiles)
Attribute inference

- We infer the following attributes:
  - **Binary**
    - Gender \{Male, Female\}
    - Relationship \{Single, Married\}
  - **Multi-value**
    - Country \{US, PH, IN, ID, GB, GR, FR, MX, IT, BR \} (top10)
    - Age group \{13-17, 18-24, 25-34, 35-44, 44-54, >54\}
Base-Line Inference

- Rely on marginal distributions
  - Maximum Likelihood of attributes
    \[ P(u.x = val \mid U) = \frac{|\{v \mid u.v = val \land v \in U\}|}{|U|} \]

- Guess the attributes’ x value from its most likely value for all users

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>51% (Male)</td>
</tr>
<tr>
<td>Relationship status</td>
<td>Unknown</td>
</tr>
<tr>
<td>Age</td>
<td>26.1% (26-34)</td>
</tr>
<tr>
<td>Country</td>
<td>23% (U.S)</td>
</tr>
</tbody>
</table>
Inference Accuracy of PubProfiles

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Baseline</th>
<th>Random guess</th>
<th>IFV Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>51%</td>
<td>50%</td>
<td>69%</td>
</tr>
<tr>
<td>Relationship</td>
<td>50%</td>
<td>50%</td>
<td>71%</td>
</tr>
<tr>
<td>Country</td>
<td>41%</td>
<td>10%</td>
<td>60%</td>
</tr>
<tr>
<td>Age</td>
<td>26%</td>
<td>16.6%</td>
<td>49%</td>
</tr>
</tbody>
</table>

TABLE IV: Inference Accuracy of PubProfiles

- More than 20% of gain in most cases
Deeper view: Gender

- It is clear from the results that music interests predict Female with a high probability.
- May be explained by the number of female profiles in our dataset (62%).

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Inferred Male</th>
<th>Inferred Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>53%</td>
<td>47%</td>
</tr>
<tr>
<td>Female</td>
<td>14%</td>
<td>86%</td>
</tr>
</tbody>
</table>

TABLE V: Confusion Matrix of Gender
Deeper view: Relationship

- It is challenging since less than 17% of crawled users disclose this attributes

- Single users are more distinguishable
  - Single users share on average 9 music interests whereas married share only 5.7

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Inferred Single</th>
<th>Inferred Married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>78%</td>
<td>22%</td>
</tr>
<tr>
<td>Married</td>
<td>36%</td>
<td>64%</td>
</tr>
</tbody>
</table>

TABLE VI: Confusion Matrix of Relationship
Deeper view: Country

- 80% of users belong to top 10 countries
- Country with specific (regional) music have better accuracy
  ➔ we clearly see the role of the semantic

<table>
<thead>
<tr>
<th>Country</th>
<th>% of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>71.9%</td>
</tr>
<tr>
<td>PH</td>
<td>7.80%</td>
</tr>
<tr>
<td>IN</td>
<td>6.21%</td>
</tr>
<tr>
<td>ID</td>
<td>5.08%</td>
</tr>
<tr>
<td>GB</td>
<td>3.62%</td>
</tr>
<tr>
<td>GR</td>
<td>2.32%</td>
</tr>
<tr>
<td>FR</td>
<td>2.12%</td>
</tr>
<tr>
<td>MX</td>
<td>0.41%</td>
</tr>
<tr>
<td>IT</td>
<td>0.40%</td>
</tr>
<tr>
<td>BR</td>
<td>0.01%</td>
</tr>
</tbody>
</table>

TABLE VII: Top 10 countries distribution in PubProfiles
Accuracy for VolunteerProfile

- The results are slightly the same as for PubProfile
- Our method is independent from the source of information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Baseline</th>
<th>Random guess</th>
<th>IFV Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>51%</td>
<td>50%</td>
<td>72.5%</td>
</tr>
<tr>
<td>Relationship</td>
<td>50%</td>
<td>50%</td>
<td>70.5%</td>
</tr>
<tr>
<td>Age</td>
<td>26%</td>
<td>16.6%</td>
<td>42%</td>
</tr>
</tbody>
</table>

TABLE IX: Inference Accuracy for VolunteerProfiles
Discussion ✓

✓ No need for frequent model updates

✓ The approach is ‘rather’ General
  ✓ OSN Independent: Many other sources of Information (deezer, lastfm, blogs, forums) etc.

✓ Use a free, open and updated encyclopedia
Discussion

- Augment the model by analyzing more interest’ category
  - Movies
  - Books
  - Sport …
- Multilanguage Wikipedia to handle foreign language
- More aggressive stemming
Conclusion

- Wikipedia Ontology to extract Semantics
- LDA to extract Topics
  - Socio, demographics, geo political aspects
  - “virtual” Communities
- K-NN to infer attributes
- The approach is general
  - Using seemingly harmless information
  - Efficient, inconspicuous profiling
If someday we all go to prison for downloading music, I just hope they split us by the music genre.
Facebook Questions
Get answers from the people you trust.
Crawling Facebook

- Crawling Facebook was challenging
  - Protection using JavaScript rendering:
    - Using a homemade lightweight browser
  - Protection using a threshold for a maximum number of request
    - Using multiple machines

- Avoiding Biased Sampling
  - Crawling Facebook public directory (100 millions users)
  - Randomly choose a user and crawl his/her profile

- Parsing HTML pages
  - It is just a mess
## Availability of attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Raw (%)</th>
<th>Pub (%)</th>
<th>Volunteer (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>79</td>
<td>84</td>
<td>96</td>
</tr>
<tr>
<td><strong>Interests</strong></td>
<td>57</td>
<td><strong>100</strong></td>
<td>62</td>
</tr>
<tr>
<td>Current City</td>
<td>23</td>
<td>29</td>
<td>48</td>
</tr>
<tr>
<td>Looking For</td>
<td>22</td>
<td>34</td>
<td>-</td>
</tr>
<tr>
<td>Home Town</td>
<td>22</td>
<td>31</td>
<td>48</td>
</tr>
<tr>
<td>Relationship</td>
<td>17</td>
<td>24</td>
<td>43</td>
</tr>
<tr>
<td>Interested In</td>
<td>16</td>
<td>26</td>
<td>-</td>
</tr>
<tr>
<td>Birth Date</td>
<td>6</td>
<td>11</td>
<td>72</td>
</tr>
<tr>
<td>Religion</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>