NDSS Symposium 2012

YOU ARE WHAT YOU LIKE

INFORMATION LEAKAGE THROUGH USERS' INTERESTS

Abdelberi (Beri) Chaabane, Gergely Acs, Mohamed Ali Kaafar

Internet = Online Social Networks ?

- Most visited websites:
  - Facebook (2sd), YouTube (3rd), Twitter (10th)

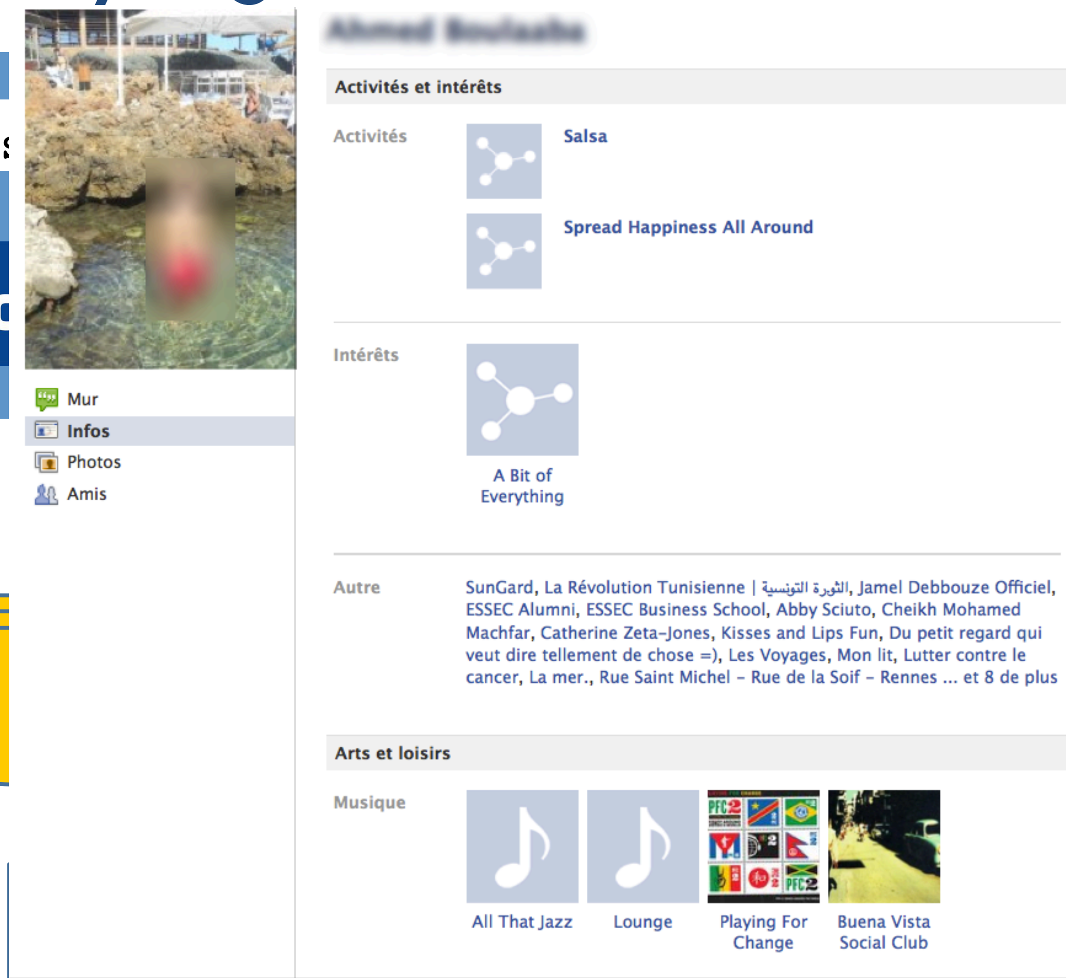
- Facebook[1]:
  - > 800M users
  - > 350M users access through their mobile
  - > 250M photos are uploaded per day
  - > 20M application installation per day

And privacy ??

1: https://www.facebook.com/press/info.php?statistics

Identifying the threat

Users



rk Z. is a

ther

setting.

.vate Profiles

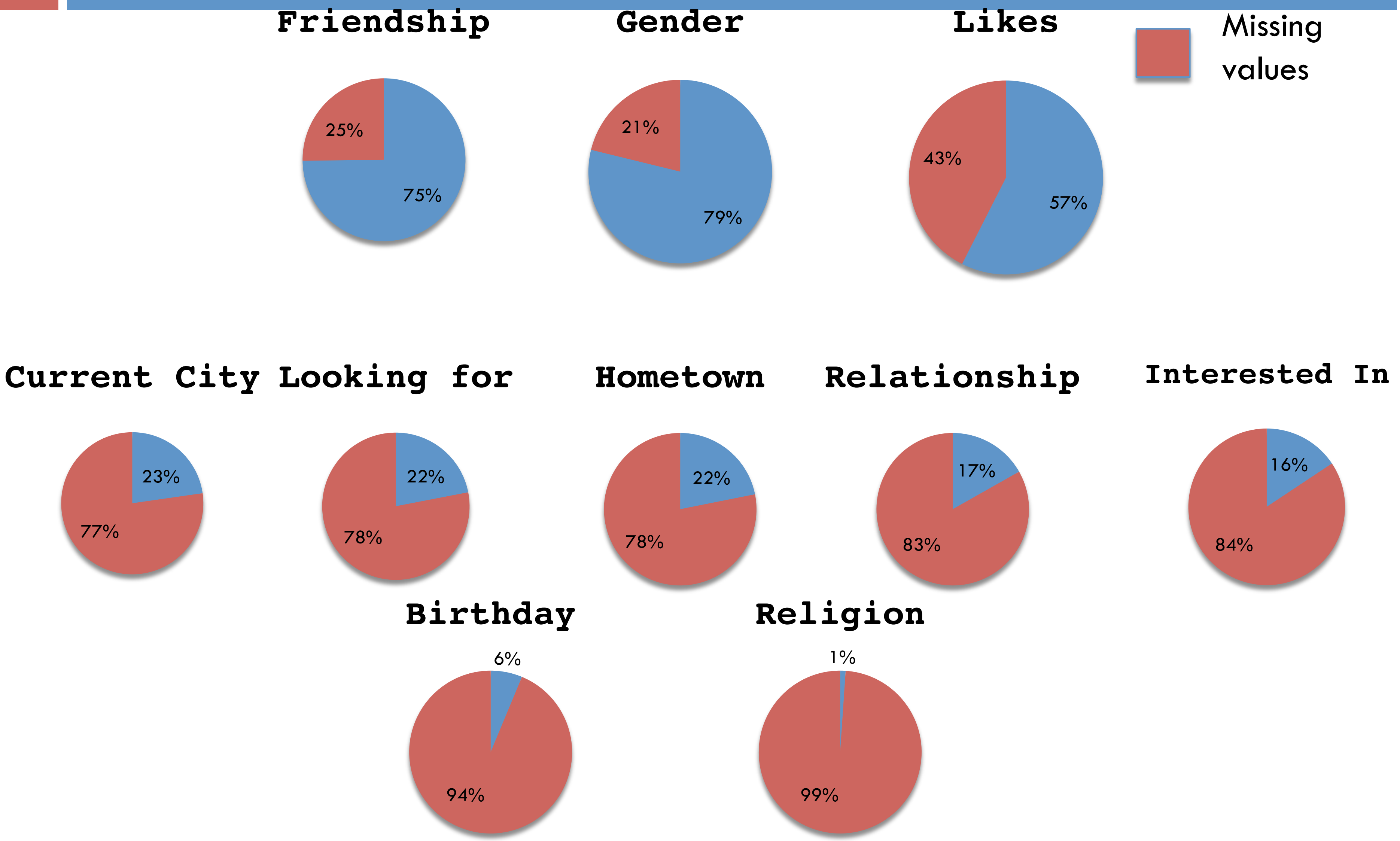
Technique

Goal

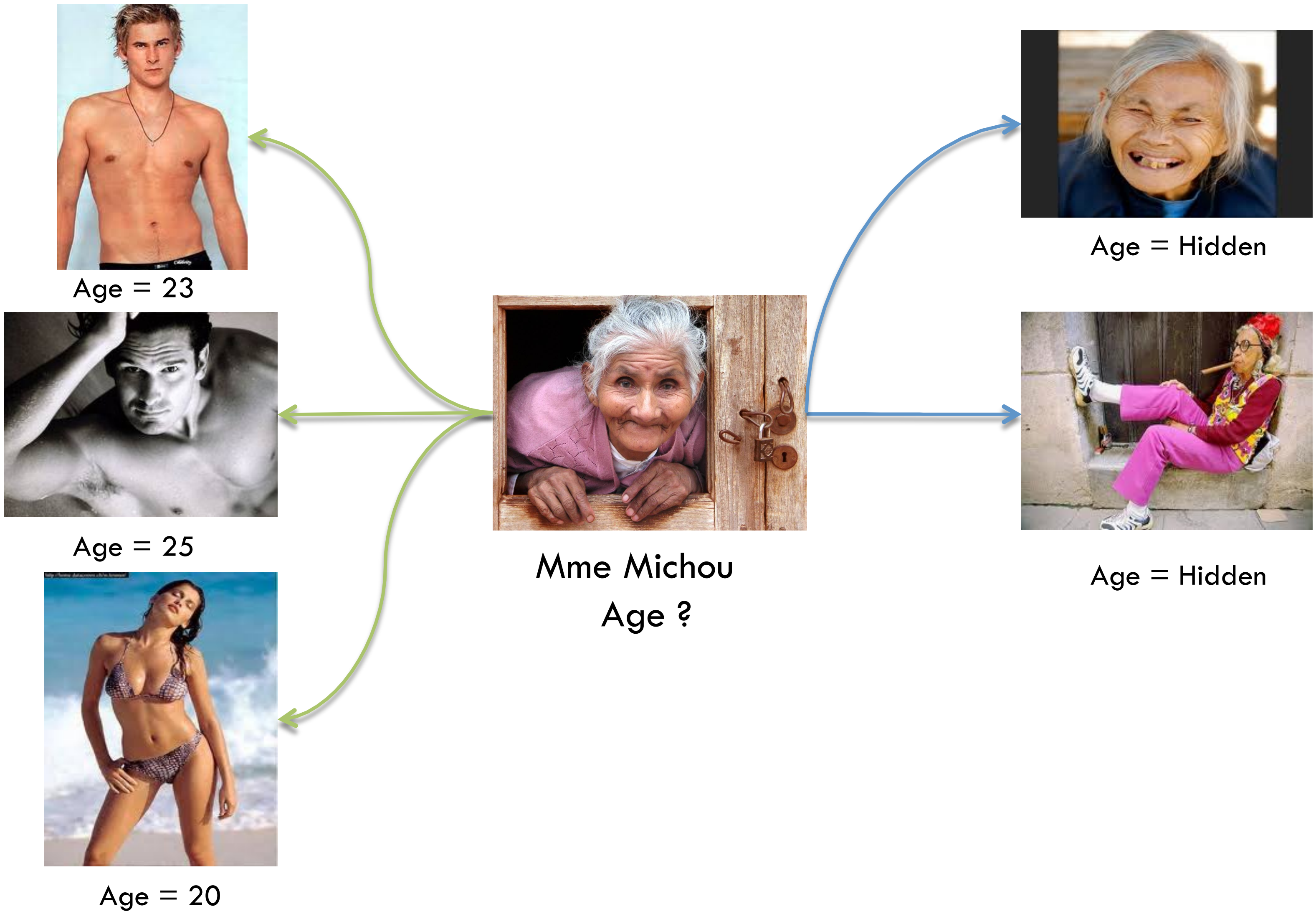
- Inferring Missing/Hidden information from a public user profile
  - ~~Using Friendship or links information~~[2,3]
  - Only using user's revealed data

*: http://13thfloorgrowingold.wordpress.com/

What people reveals ?



Homophily or not homophily



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 ?

likes

- ✓ 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

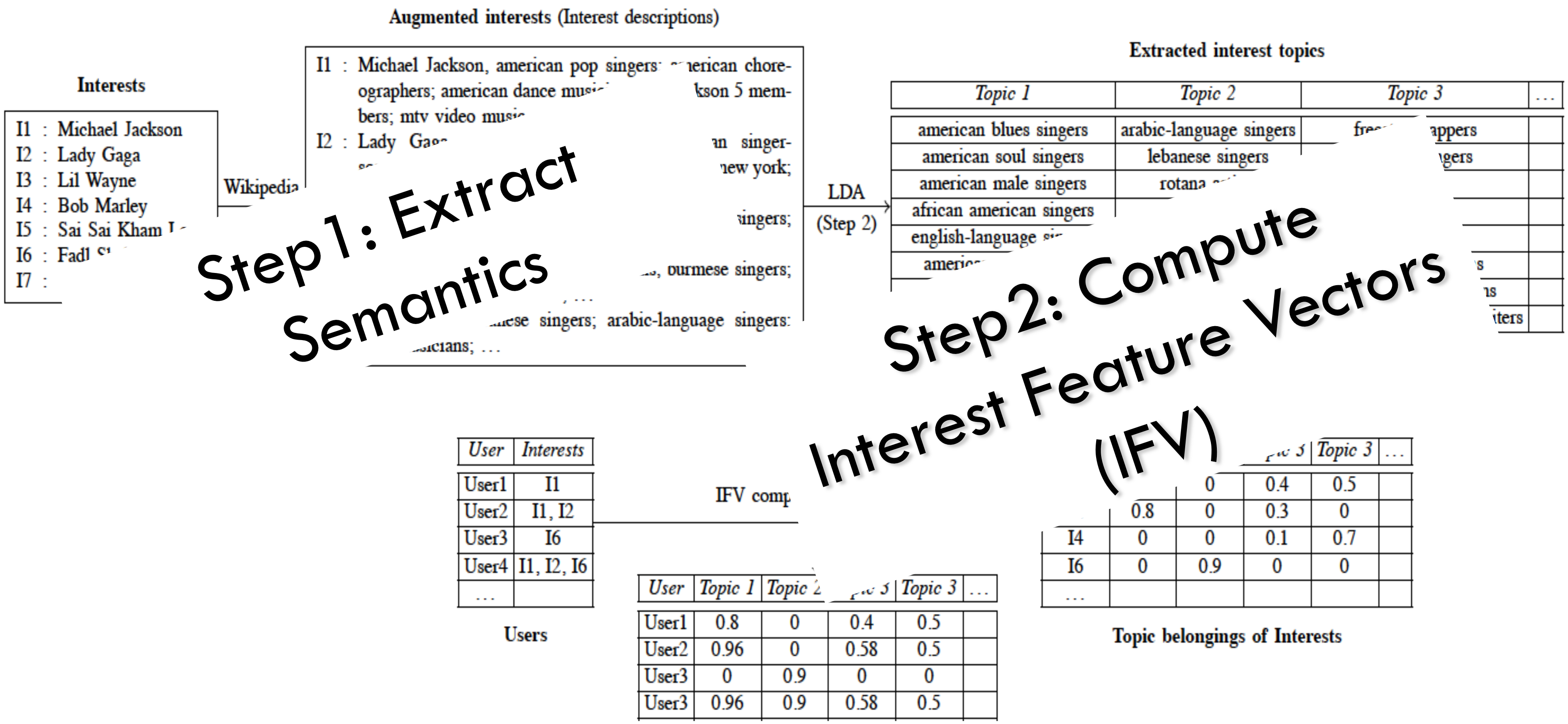
Btw it belongs to
http://facebook.com/kave.salamatian

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

Topic about Iran	Topic about Islam (Religion)	Topic about classical music
Iranian classical Vocalist Iranian humanitarians Iranian Iranian Kurdish people people of Azerbaijani Persian-language …	Iranian Shi'a Muslims People	vocalists Iranian Iranian classical vocalists

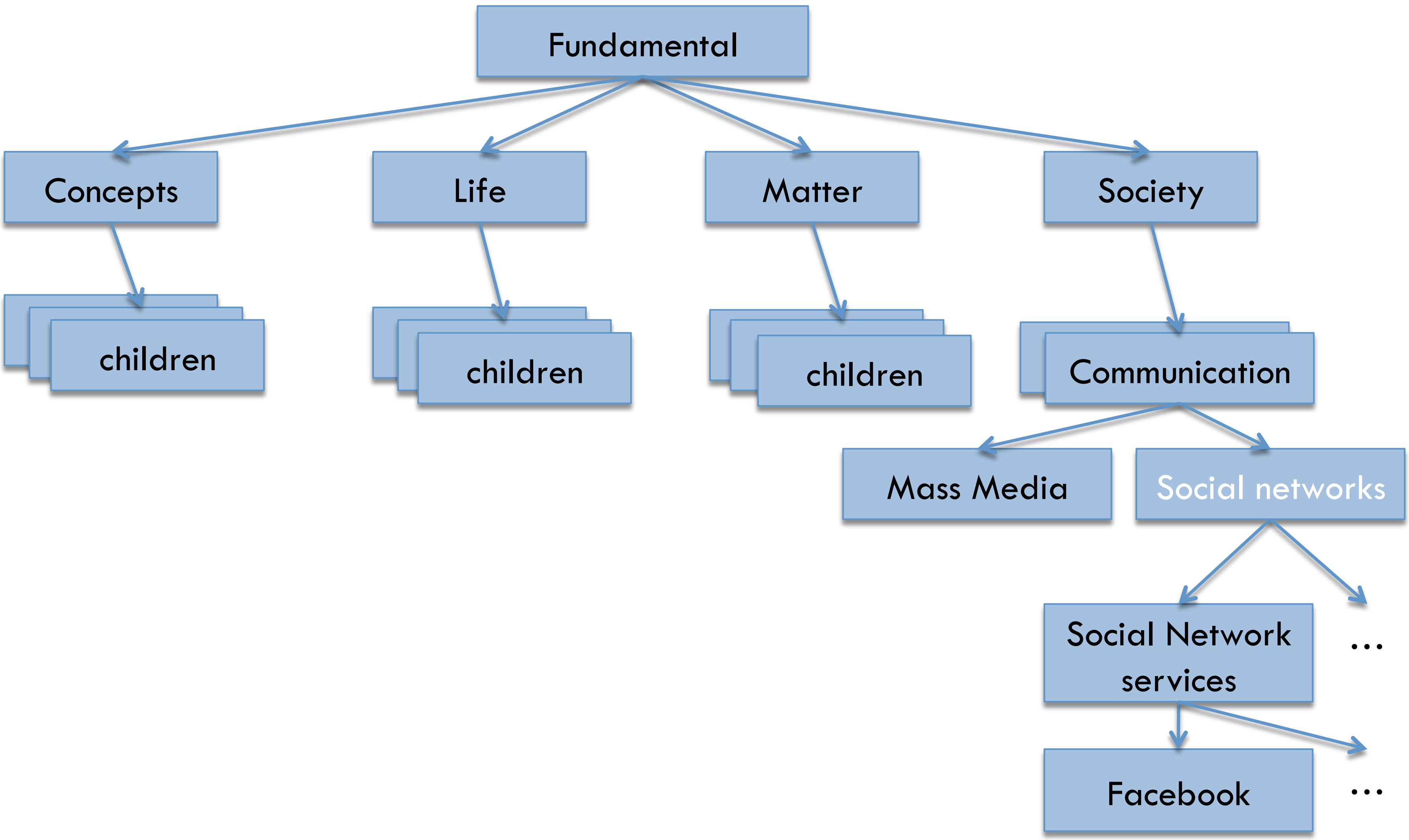
The Algorithm



Step3: Classify Users
Infer Attributes

$$Step 1$$

Tree of wikipedia



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

Extract semantic (Description)

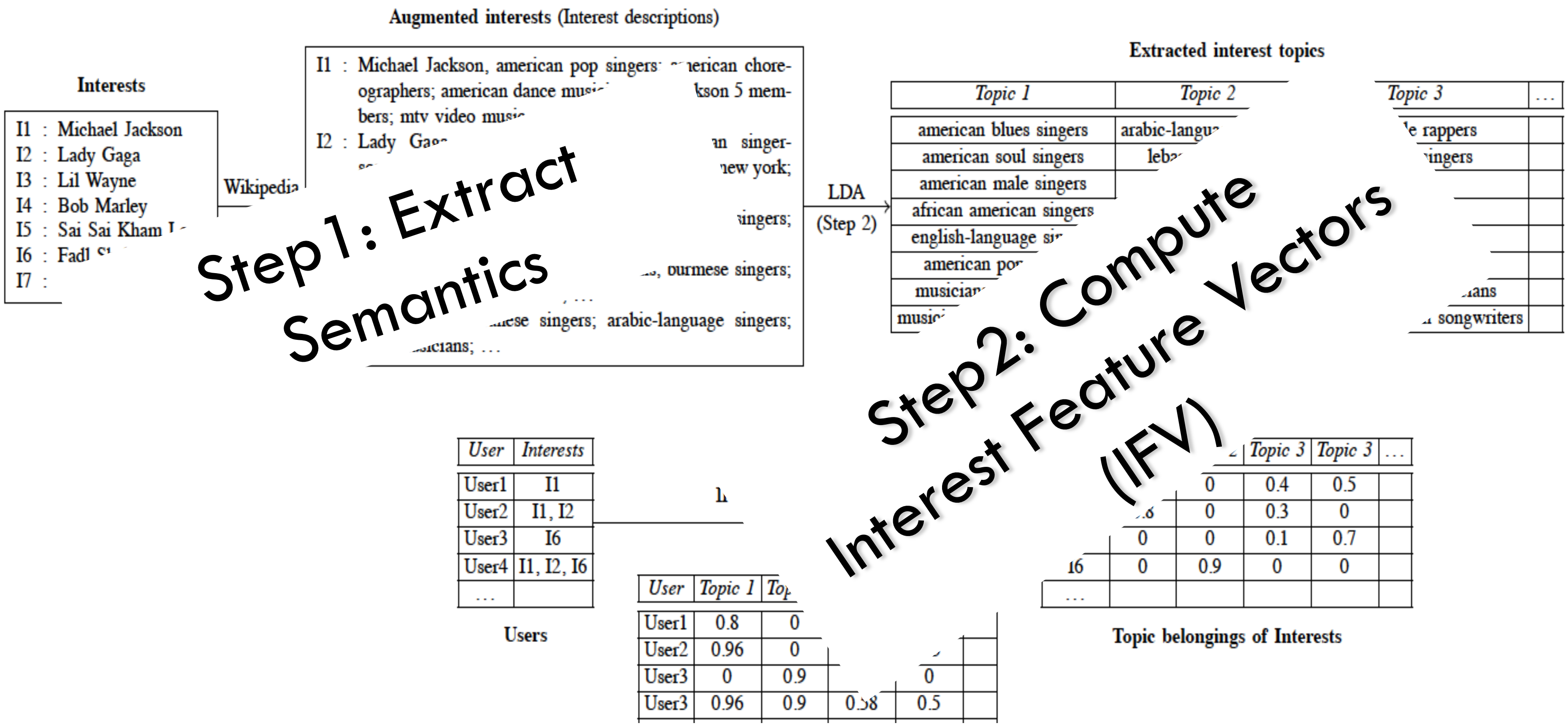
- Using the same granularity allows us to semantically 'link' similar concepts

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

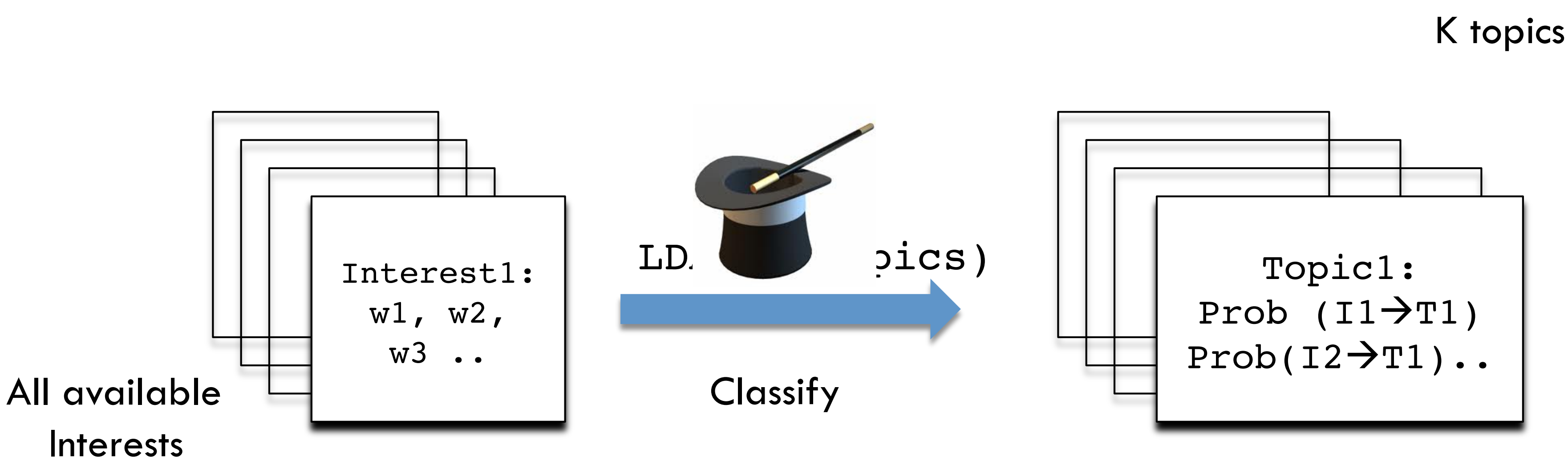


Step3: Classify Users

Infer Attributes

$$Step2$$

LDA Intuition



I1: Interest1
T1: 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 ?

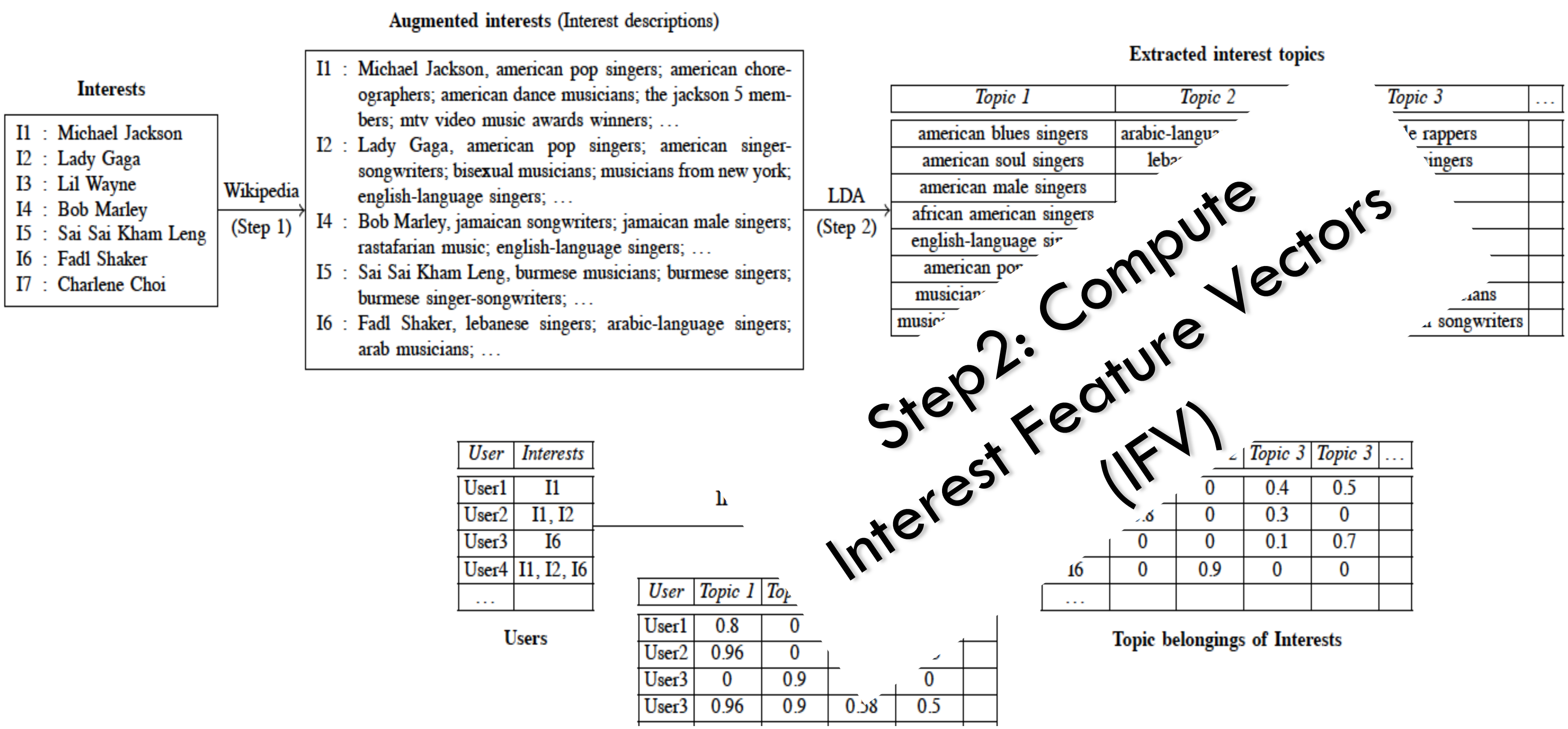
D.Blei (MLSS'09)

LDA

- 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 documents to belong to each topic

Topic example

The Algorithm



Step3: Classify Users

Infer Attributes

$$Step3$$

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 $O(M \log_2 M))$

Example

	IFV				Attribute to infer
user1	0.2	0.6	0.1	0.1	Attribute=?
~~user2~~	~~0.3~~	~~0.6~~	~~0.7~~	~~0.01~~	~~Attribute=?~~
user3	0.2	0.4	0.1	0.1	Attribute=Val
…					
User *n*	0.1	0.1	0.1	0.1	Attribute=Val

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

The **n** nearest users to user1 are: S={user3, user*m*, …}

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

$$P(u.x = val \mid U) = \frac{|\{v \mid u.v = val \wedge v \in U\}|}{|U|}$$

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

Attribute	Value
Gender	51% (Male)
Relationship status	Unknown
Age	26.1% (26-34)
Country	23% (U.S)

Inference Accuracy of PubProfiles

Attribute	Baseline	Random guess	IFV Inference
Gender	51%	50%	69%
Relationship	50%	50%	71%
Country	41%	10%	60%
Age	26%	16.6%	49%

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%)

Attribute \ Inferred	Male	Female
Male	53%	47%
Female	14%	86%

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

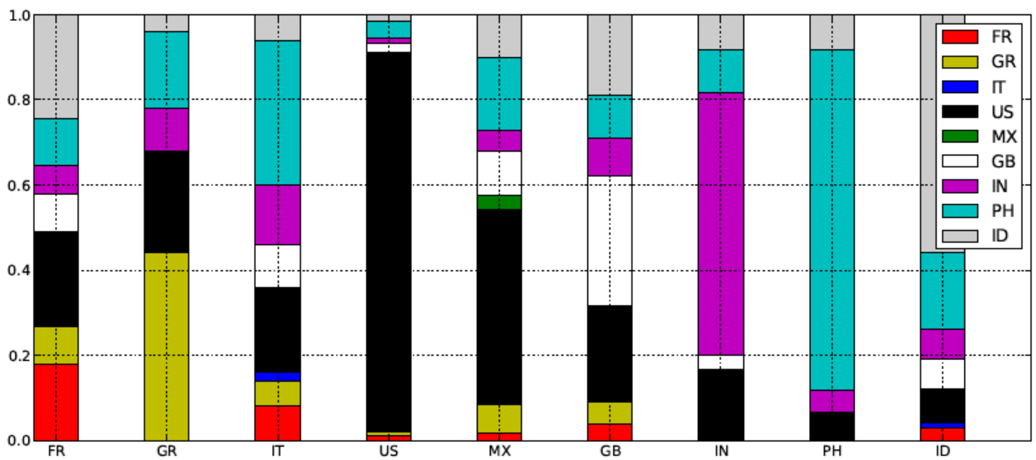
Attribute \ Inferred	Single	Married
Single	78%	22%
Married	36%	64%

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**



Country	% of users
US	71.9%
PH	7.80%
IN	6.21%
ID	5.08%
GB	3.62%
GR	2.32%
FR	2.12%
MX	0.41%
IT	0.40%
BR	0.01%

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

Attribute	Baseline	Random guess	IFV Inference
Gender	51%	50%	72.5%
Relationship	50%	50%	70.5%
Age	26%	16.6%	42%

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 ✗

Dislike

- 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.
THE BEST FUN SITE = 9GAG.COM

~~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

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