AUDIO ERGO SUM

A Peronal Data Model For Musical Preferences

Riccardo Guidotti
Giulio Rossetti
Dino Pedreschi

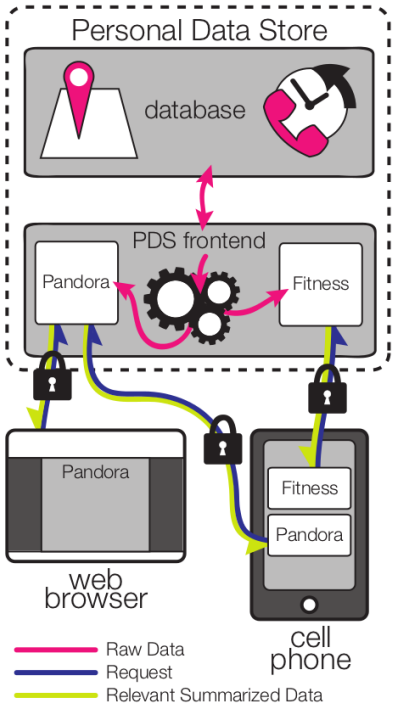
Why user centric solutions?

- We produce an unthinkable amount of data while running our daily activities.
- How can we manage all these data?
- Can we get an added value from them?

Related Works: openPDS

- Collect, store, and give fine grained access of metadata to third parties.
- Protects the privacy of metadata by applying a hard anonymization.
- It allows services to ask questions whose answers are calculated against the metadata.
- High dimensional data to low dimensional answers less likely to contain sensitive information.
- Answers can be shared individually or in aggregate.

openPDS: Protecting the Privacy of Metadata through SafeAnswers, Yves-Alexandre de Montjoye, Alex Sandy Pentland, et al. PlosONE, 2014



Related Works: PIMS

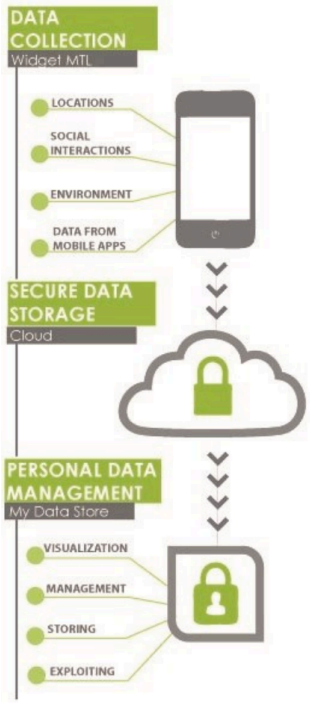
Managing Your Digital Life, Serge Abiteboul, Benjamin André, and Daniel Kaplan, Commun, 2015

- It consists of a user's server, running the services selected by the user, storing and processing the user's data.
- The user pays for the server so the server does what the user wants it to do.
- The user chooses the application code to deploy on the server.
- The server software is possibly open source.
- The server resides in the cloud so it can be reached from anywhere.

Related Works: MyDataStore

- It enables to control and share data organized as a set of web-based services.
- *Collection*: users can determine which data automatically collect and store.
- *Sharing*: users can choose whether to disclose or not their data and in which detail (e.g. anonymously).
- *Deletion*: users can delete single records or all data collected in a specific region and time interval.
- Views with different levels of aggregation: (i) *individual* increasing user's consciousness; (ii) *social* using data shared by.

My Data Store: Toward User Awareness and Control on Personal Data, Michele Vescovi et all. Ubicomp, 2014



Data Type

- A listening ***l=<timestamp,song,artist,album,genre>*** is a tuple formed by the timestamp indicating when the listening occurred, the song listened, the artist which plays the song, the album the song belongs to, and the genre of the artist.

- Given a user ***u*** we define with ***L_u = { l_i }*** the set of listening performed by ***u***.

Model Extraction

- From the listening L_u we can extract her Personal Listening Data Model P_u by employing data mining techniques.

Sets

- From the listening L_u we can extract the set of songs S_u, artists A_u, albums B_u, and genres G_u.
- **The sizes of these sets are a first simple and valuable indicators.**

- Example

$A_u = \{The\ Beatles, Muse, Shakira, \ldots\}, |A_u| = 52$

$G_u = \{Rock, Folk, Pop, Metal, \ldots\}, |G_u| = 34$

Support

- The *support dictionary* of a certain *feature* is a set of couple *(item, support)* where the support of an item is relative number of occuring items.
- Besides ***S_u***, ***A_u***, ***B_u*** and ***G_u*** we define the support dictionary also **considering the day-of-the-week *D***, and **the time-of-the-day *T***.

- Example

```
a_u = {(Muse, 0.12),(The Beatles, 0.23), …}
g_u = {(Rock,0.3),(Folk, 0.22), …}
```

Entropy

- The normalized Shannon Entropy tends to 0 when the user behavior is systematic, tends to 1 when the behavior is not predictable.

$$entropy(X) = \frac{-\sum_{i=1}^{n} \mathrm{P}(y_i) \log_2 \mathrm{P}(y_i)}{\log_2 n} \quad \in [0, 1]$$

- Example

```
g_u = {(Rock,0.8),(Folk, 0.1), (Pop, 0.1)}
e_gu = 0.58

g_v = {(Rock,0.4),(Folk, 0.3), (Pop, 0.3)}
e_gv = 0.99
```

Top & Repr

- **Top:** is the **item** with the **highest support** with respect to a certain feature.

$$top(X) = \underset{(x,y)\in X}{argmax}(y)$$

- **Repr:** is the **set of items significantly most listened**, i.e. with the highest supports, with respect to a certain feature.

$$repr(X) = \underset{(x,y)\in X}{knee}(y) = \underset{(x,y)\in X^*, y'\in X'}{argmax}(|y - y'|)$$

- Example

```
g_u = {(Rock,0.4),(Pop,0.3),(Folk,0.1),
(Classic,0.1),(House,0.1)}

top(g_u) = (Rock,0.4)
repr(g_u) = {(Rock,0.4),(Pop,0.3)}
```

Listening Sequences

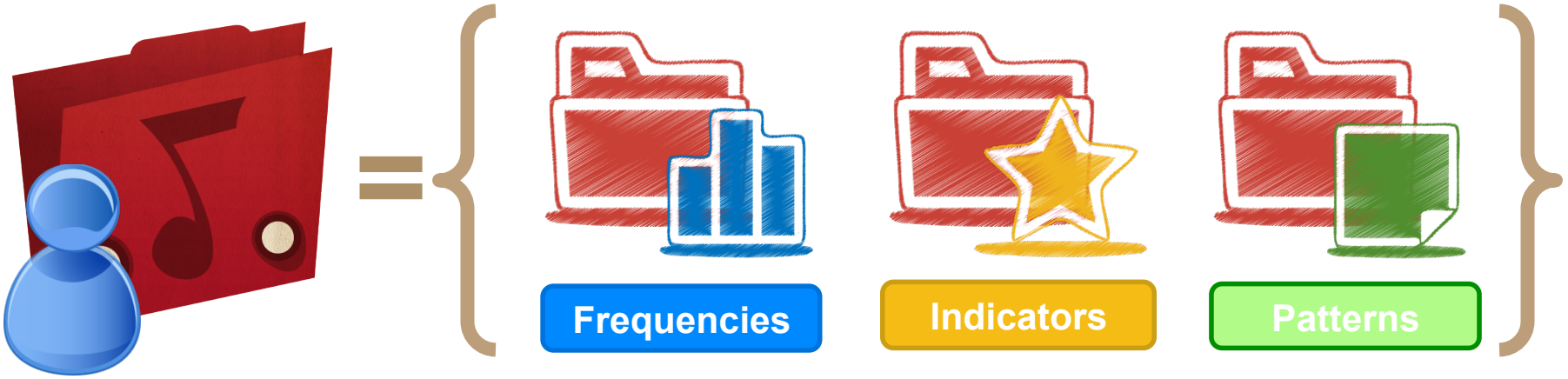
- A sequence is a list built by concatenating the items of the listening L_u in a given time window, orederd by timestamp and describing a feature

- **A frequent sequence** is a closed frequent sequence with at list **minsup occurrences.**

- Example
- *L_u = {<t0,.,Rock,.,.>, <t1,.,Rock,.,.>, <t2,.,Folk,.,.>, <t3,.,Classic,.,.>…, <t10,.,Rock,.,.>, <t21,.,Rock,.,.>, <t22,.,Folk,.,.>, <t3,.,Pop,.,.>…}*
- *Seq = {[Rock,Rock,Folk,Classic], [Rock,Rock,Folk,Pop]}*
- *FG_u = {([Rock,Rock,Folk], 0.10), …}*

Personal Listening Data Model

- The PLDM characterizes the listening behavior of a user by means of its:
  - **indicators**: set sizes and entropy values;
  - **frequencies**: support dictionaries;
  - **patterns**: most listened, most representative and frequent sequences.

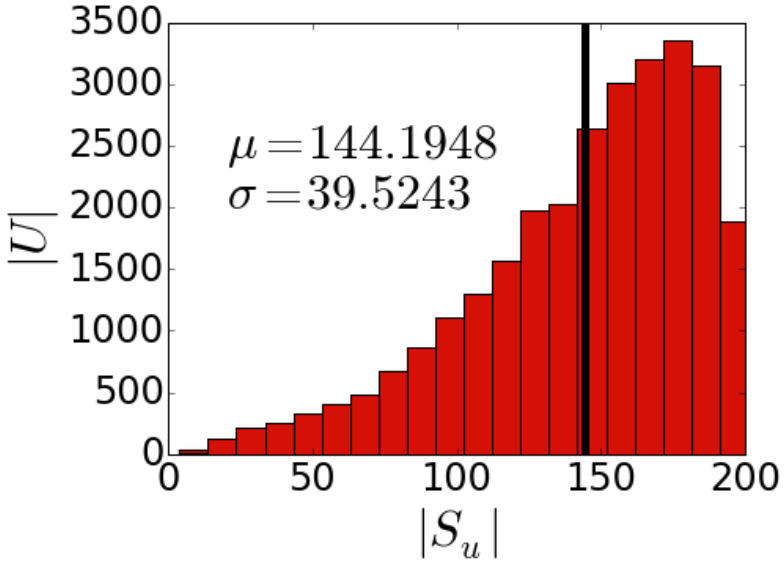


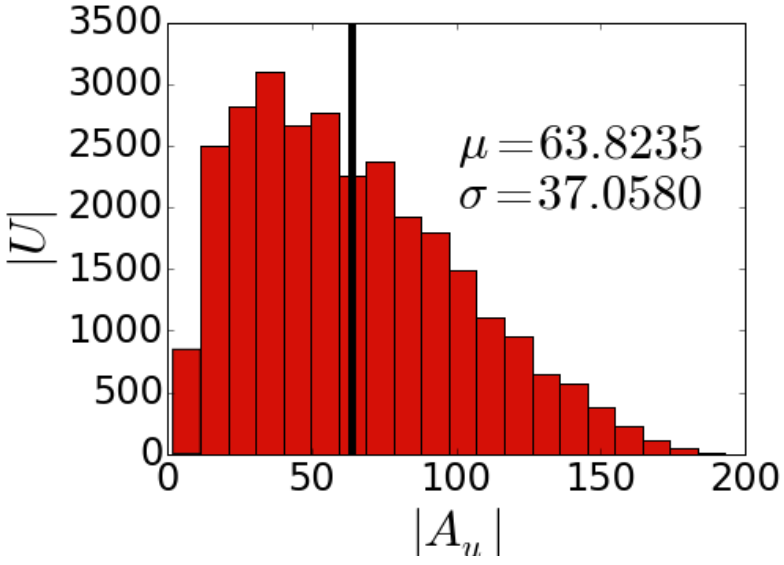
LastFM

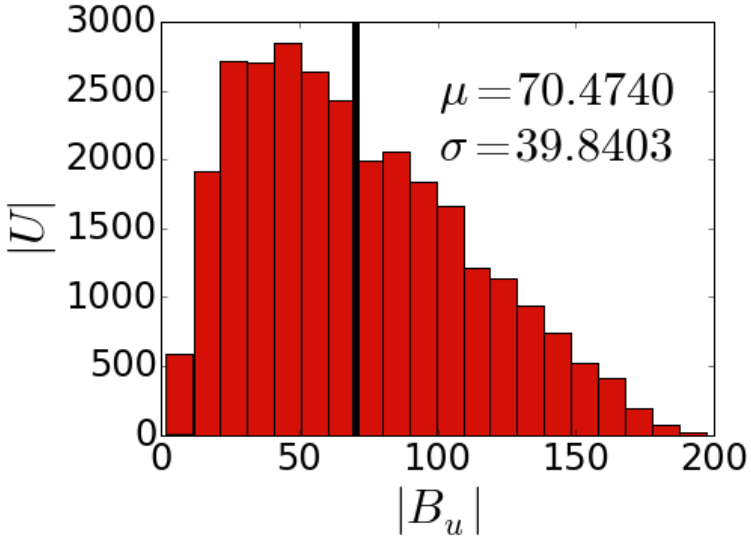
- ***Last.FM*** is an online social network where people can share their own music tastes and discover new artists and geners on the bases of what they, or their friends, like.

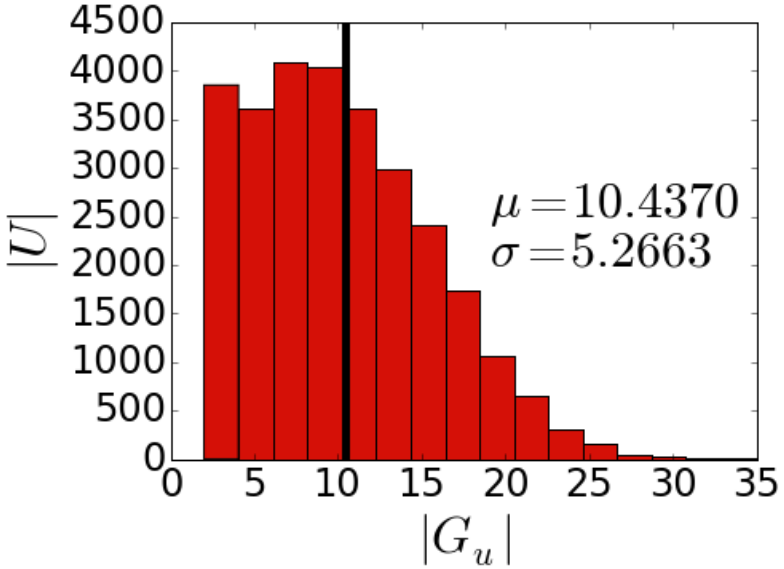
- ***Dataset***
  - Users: 30,000
  - Region: UK
  - Listenings: 6,000,000

PLDM Indicators Analysis – Set Size

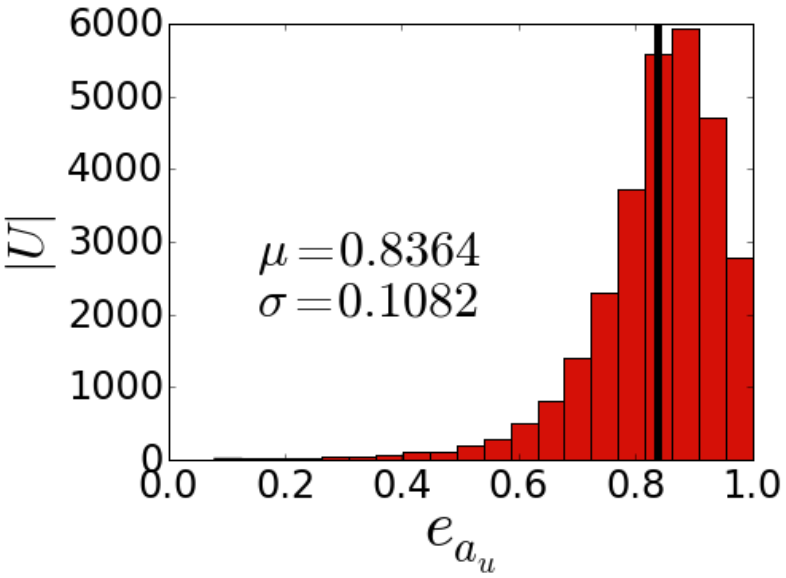


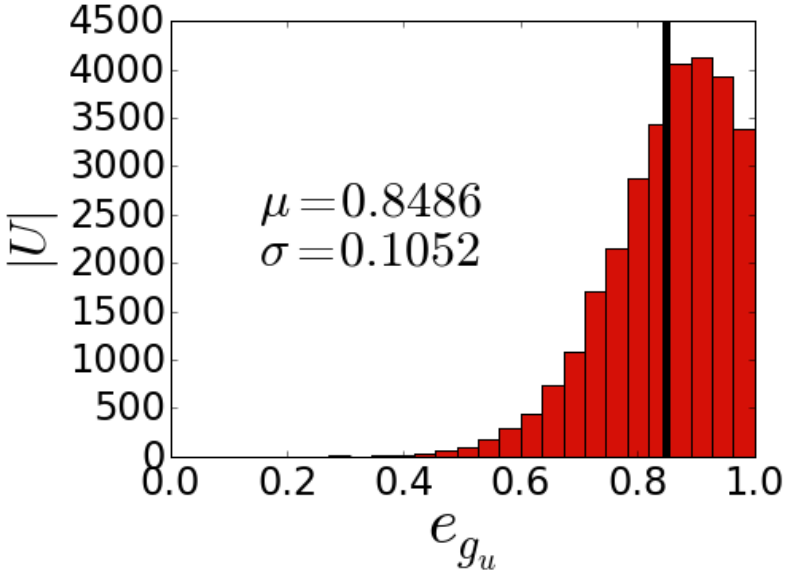


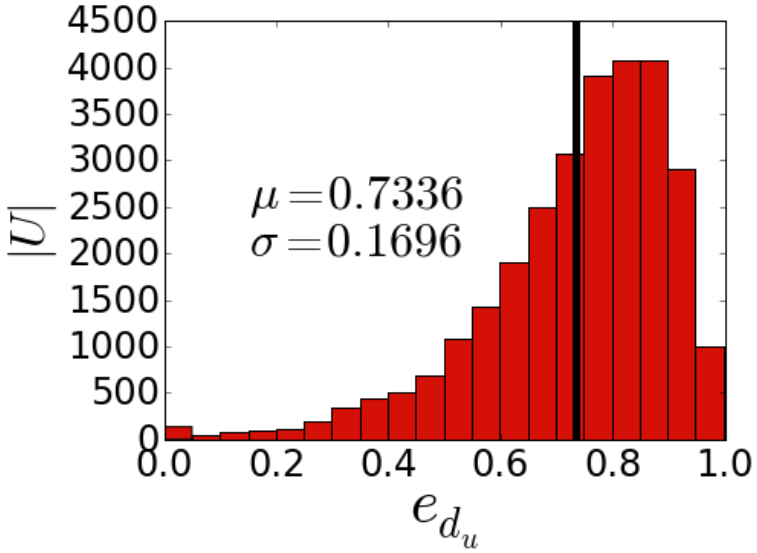


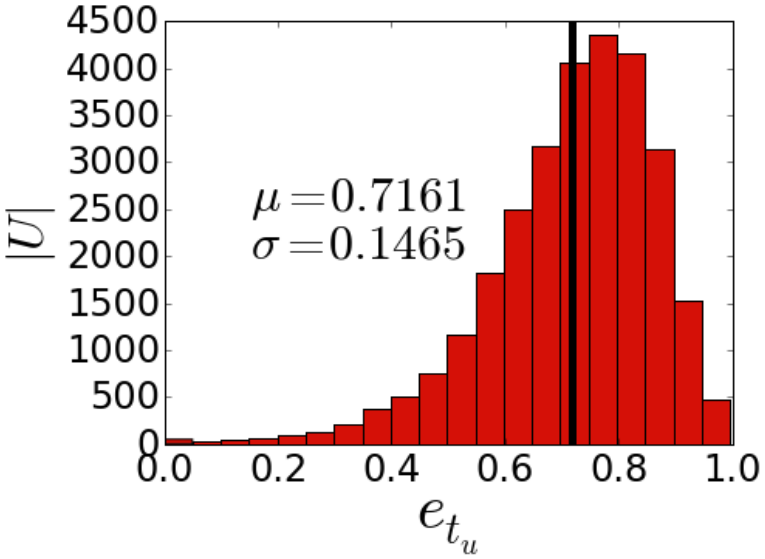


PLDM Indicators Analysis – Entropy

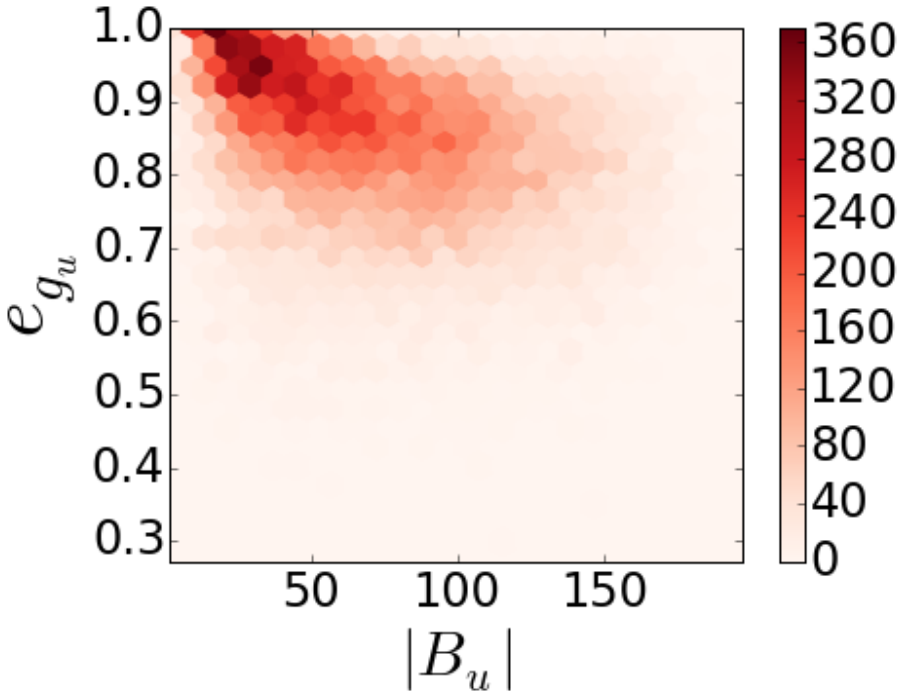


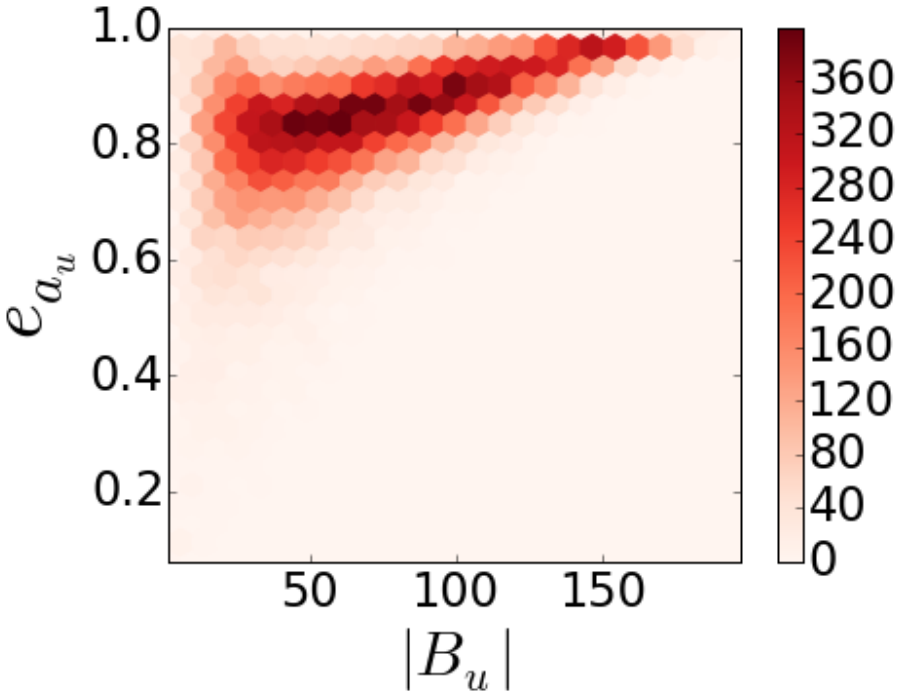






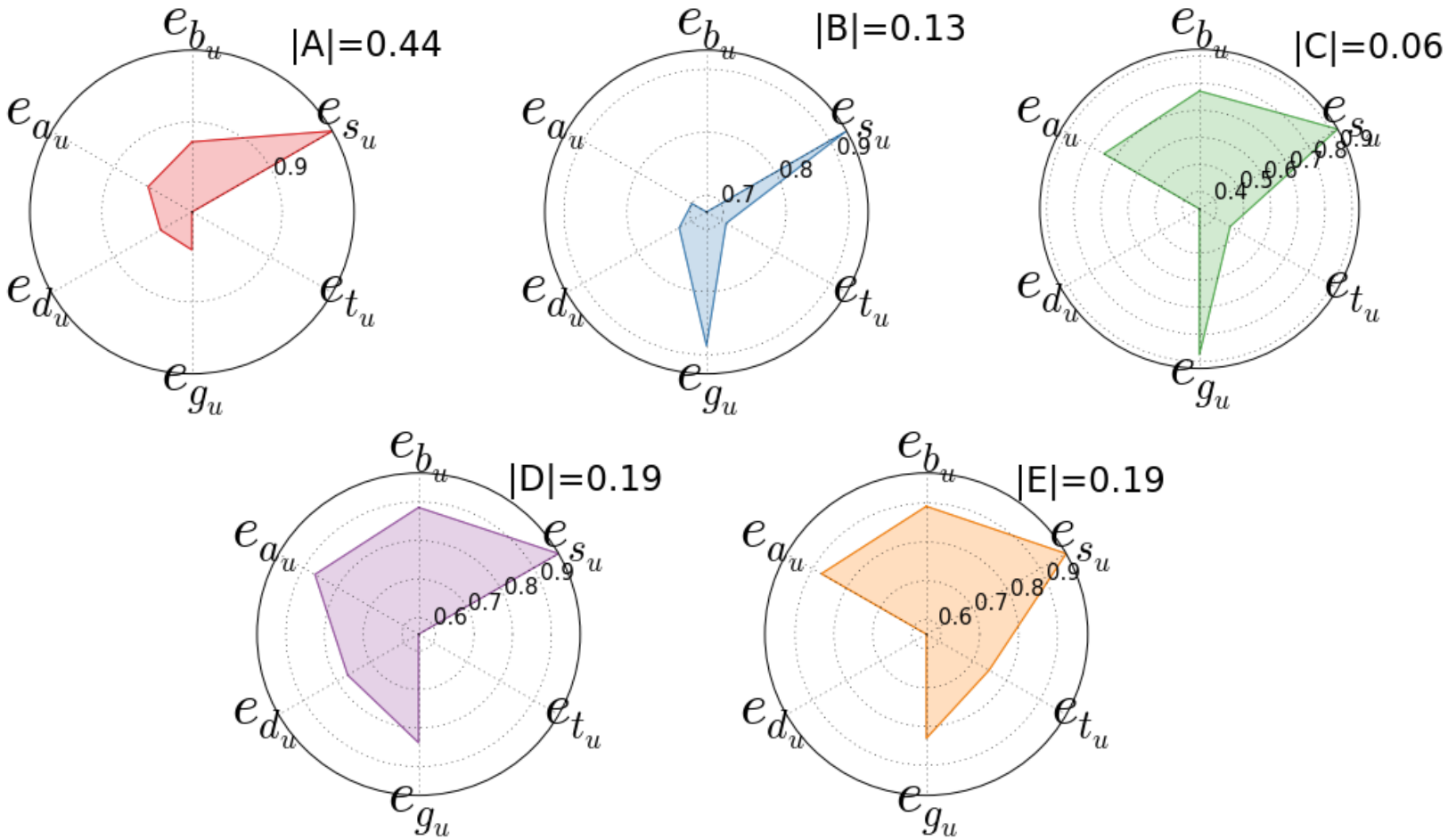
PLDM Indicators Analysis – Entropy



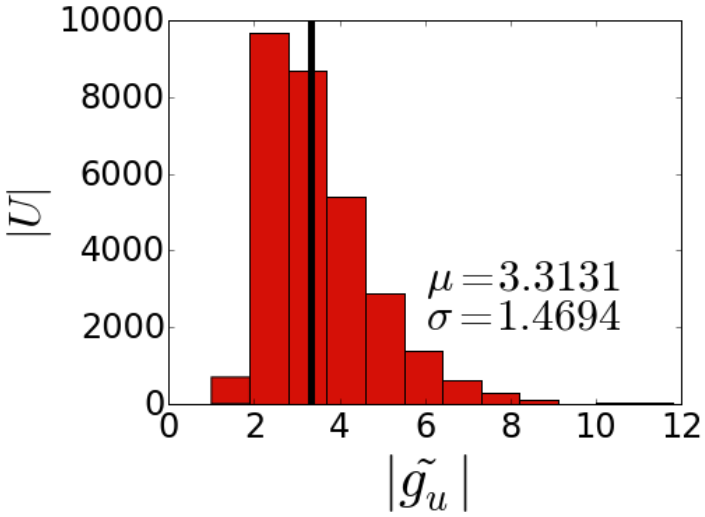


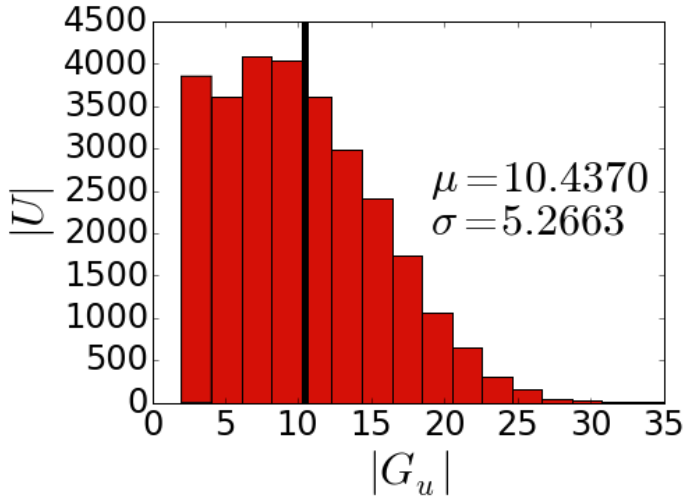
PLDM Segmentation Analysis

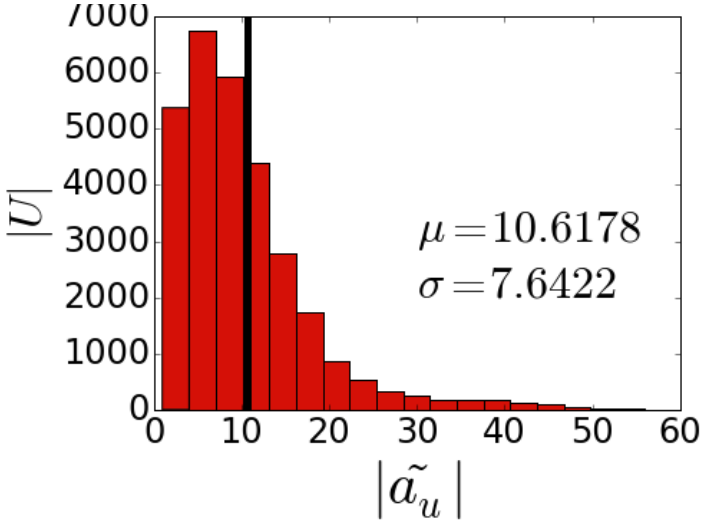
- We applied K-Means to the indicators with k=5.

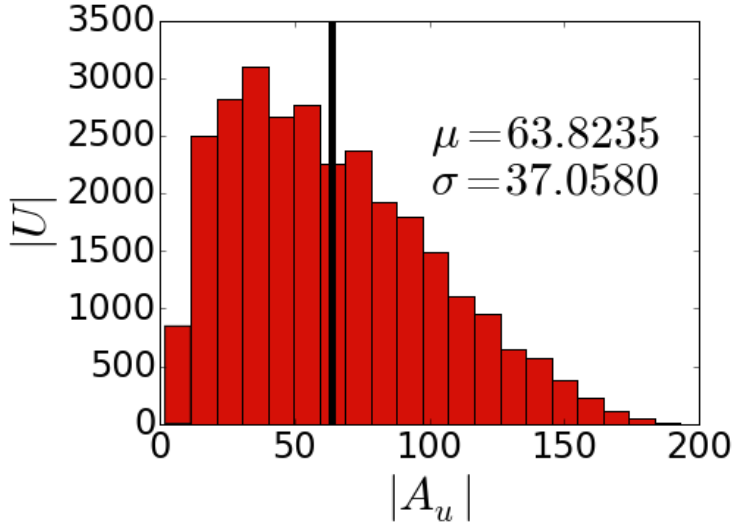


PLDM Frequency Analysis - Size









Conclusion & Future Work

- We propose a ***personal data model*** for analyzing and managing listening behavior.
- The PLDM can be applied both for ***individual*** and ***collective analysis*** and ***services*** (e.g. recommendations).
- We discovered that we are characterized by a limited set of musical preferences, but not by a unique predilection.

- Add to the model the friendship dimension and analyze its impact to estimate users homophily with respect to listening.
- Implement a real web service dashboard for self-awareness able to visualize the patterns and the indicators of the PLDM.

THANK YOU

riccardo.guidotti@di.unipi.it