30Music listening and playlist dataset

R. Turrin, M. Quadrana,
A. Condorelli, R. Pagano, P. Cremonesi

Challenges and peculiarities

		MSD	Celma 1K	Celma 360K	Yahoo! R1	Yahoo! R2	Yahoo! C15	artofthemix	#nowplaying	MMTD	Music Micro	30Music
	Sources	(a)	(b)		(c)			(d)	(e)	(f)		(b)
Features	Release	2011	2010	2010	n/a	n/a	n/a	2003	live	2013	2012	2015
	Span	n/a	4y	n/a	1m	5y	10y	1m	4.5y	18m	1y	1y
	Users	??	1K	360K	2M	1.8M	-	1M	5M	15K	136K	45K
	Tracks	1M	1M	-	-	136K	625K	218K	764K	134K	71K	4.6M
	Artists	44K	180K	160K	9.4K	100K	-	60K	95K	25K	19K	600K
	Playlists	-	-	-	-	-	-	29K	-	-	-	280K
	Track info	x	-	-	-	x	-	-	x	-	-	x
	Acoustic	x	-	-	-	-	-	-	-	-	-	-
	User info	x	x	x	-	-	-	-	-	-	-	x
Interactions	Play events	-	20M	-	-	-	-	-	63M	1M	600K	31M
	Playtime	-	-	-	-	-	-	-	-	-	-	x
	Sequences	-	-	-	-	-	-	-	-	-	-	2.7M
	Ratings	-	-	-	12M	717M	262M	-	-	-	-	1.7M
	Play count	??	-	17M	-	-	-	-	-	-	-	-

(a) EchoNest, MusicBrainz, LastFM (b) LastFM (c) Yahoo! (d) artofthemix (e) Twitter (f) Twitter, MusicBrainz, Yahoo! Place Finder, Last.fm, 7digital, allmusic.com

There is **no other publicly available dataset** which combines **sessions, user explicit preference, playlists** and **timestamps** all together.

This dataset can be used to evaluate:

- **SARS: Session Aware** Recommender Systems
- **Music** Recommender System with **explicit** and **implicit feedback**
- **Playlist** Recommendation
- **Tag** Recommendation

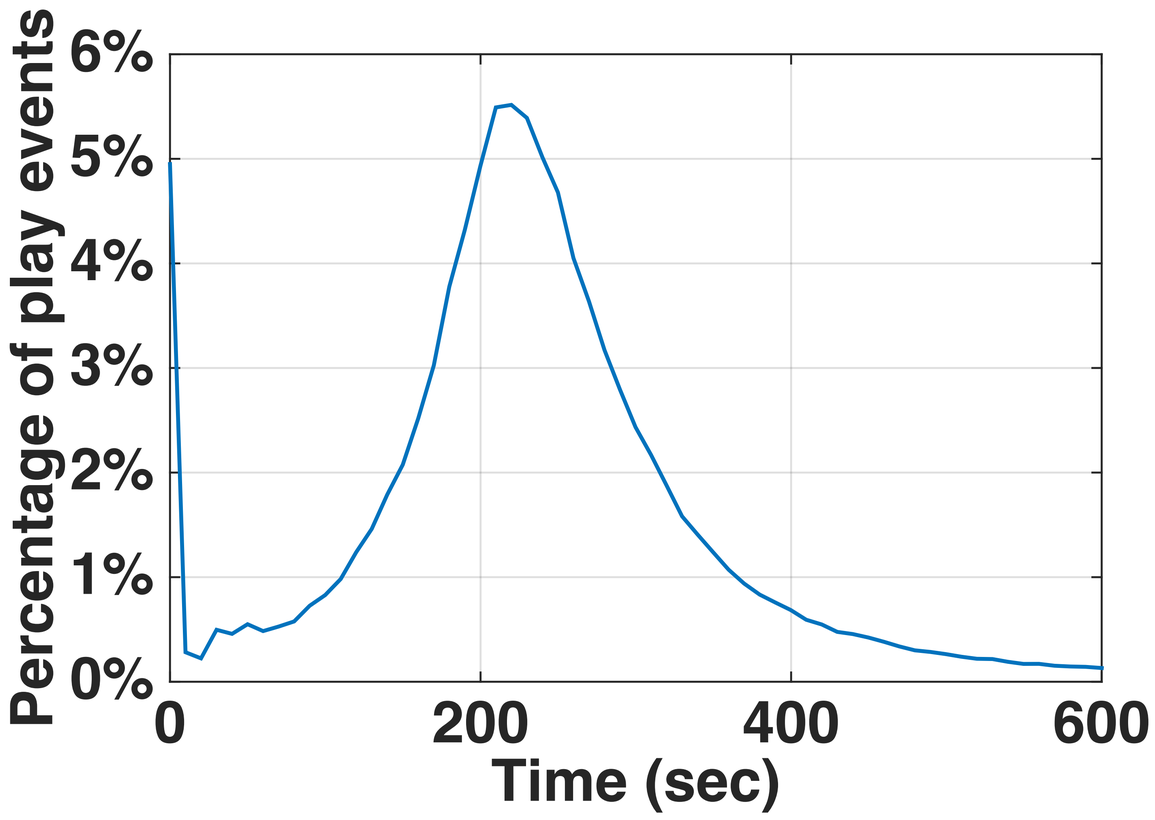
Dataset collection and session generation

1 Start from a list of 2M usernames of Last.fm users

2 Get Listening events for one year period **from 20 Jan 2014 to 20 Jan 2015**

3 Clean (e.g., discard users with no playlists)

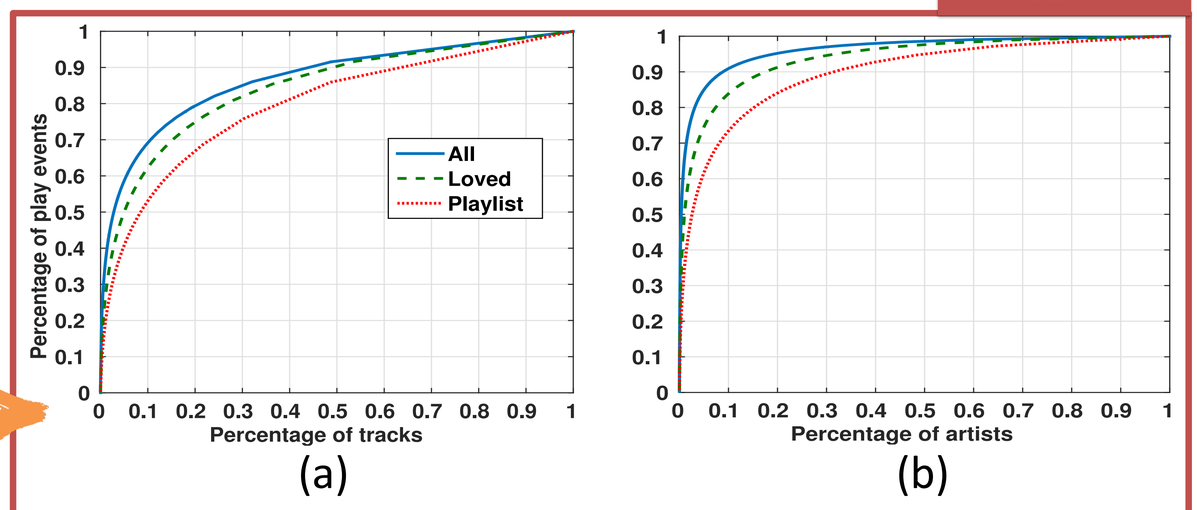
4 Merge "consecutive" play events into sessions based on a 800s threshold



Probability density function of **play durations**:

- Many tracks are **skipped** within the first 5 seconds
- Almost 100% of tracks are listened for **less than 800 s**

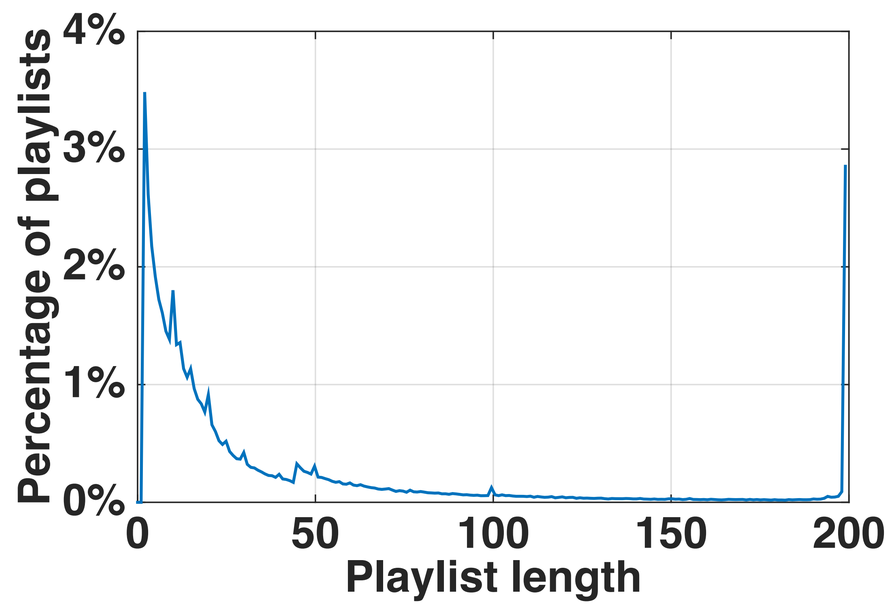
Statistics



(a) (b)

Cumulative distribution function of play events for tracks (a) and artists (b).

Listening behavior is **well distributed among the tracks**, while **artists present a stronger short-head / long tail distribution**



Probability density function of **playlist length**:

- Many playlist are composed by **only one track**
- Many other are composed by **200 tracks**: this is the maximum playlist length imposed by LastFM

PAPER PDF Scan this code or go to http://goo.gl/PHBb7r

POSTER PDF Scan this code or go to http://goo.gl/fy6vUJ

DATASET Scan this code or go to http://goo.gl/dm8gJl

The research leading to these results was performed in the CrowdRec project, which has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement n.610594.