# COS 513 Presentation 3

Danny, William, Dakota

September 30, 2015

## 1 Project Description

We are studying the types of users who use last.fm, and the songs as well. Our goal is to find the right model for users and songs and learn something new, eventually taking into account the temporal component of the play data.

## 2 Dataset

We are using the last.fm plays dataset.
`http://www.dtic.upf.edu/ ocelma/MusicRecommendationDataset/lastfm-1K.html`
Total Lines: 19,150,868
Unique Users: 992
Artists with MBID: 107,528
Artists without MBDID: 69,420

The dataset contains song data in the form

`userid \t timestamp \t musicbrainz-artist-id \t artist-name \t musicbrainz-track-id \t track-name`

and user data in the form

`userid \t gender ('m'|'f'|empty) \t age (int|empty) \t country (str|empty) \t signup (date|empty)`

## 3 Initial Data Analysis

The first thing to do is cluster the songs without taking into account the temporal component of the data. We use only the play data for this clustering. To do this, we first crawl through the data and construct a sparse matrix $M$ such that $M(i, j)$ is the number of times user $i$ played song $j$. This matrix has 1486727264 entries, 0.3% of which are nonzero. Using this matrix, we hope to cluster the songs and users. Similarly to the Netflix problem, we assume low rank, which translates to there being a few archetypical users which everyone else is a combination of, where an archetypical user is represented by a playlist. This makes PCA a natural choice for clustering. Note that we cannot center the data because doing so would destroy sparsity. Therefore, we expect the first principal component to be uninformative of clustering.

## PCA

Our goal in this exploratory analysis is to see if the play matrix carries enough information to cluster songs in some reasonable way. Our metric for success is the eye test. Since the play matrix does not understand genre, or use the artist directly as a feature, we hope that PCA will recover genre and/or artist in the playlist of a *typical* user. We choose to use the log of the number of plays instead of the number of plays itself. This is because we don't want to let outliers control our results, and using logs was a good way of differentiating between no plays, some plays, and tons of plays. As an added convenience, it removes songs that were played only once, which we don't want to weight very heavily.

PCA works reasonably well. Here are the top 20 songs in select principal components.

```
Component Number:  1
29242 Metallica: Nothing Else Matters 0.039512583237
40550 Metallica: Master Of Puppets 0.0390164041261
40552 Metallica: One 0.0368443433159
32911 Metallica: Enter Sandman 0.0367895312332
10941 Dread Zeppelin: Stairway To Heaven 0.0365366856085
33470 Metallica: Sad But True 0.0329327752888
33455 Metallica: The Unforgiven 0.0327713739416
37227 Metallica: Fade To Black 0.0325640601586
12321 Guns N' Roses: Welcome To The Jungle 0.0315305562435
38146 Metallica: Battery 0.0314776978129
63811 Queen: Bohemian Rhapsody 0.0301365930856
38119 Metallica: Wherever I May Roam 0.0301281480811
18800 Guns N' Roses: November Rain 0.0301080962211
12132 Black Sabbath: Paranoid 0.0295927419139
38117 Metallica: For Whom The Bell Tolls 0.0294124525187
32969 System Of A Down: B.Y.O.B. 0.0287993789382
42995 Motörhead: Ace Of Spades 0.0285907857447
40569 Guns N' Roses: Paradise City 0.0285214956135
32244 Pink Floyd: Comfortably Numb 0.0284861119442
61775 Guns N' Roses: Sweet Child O' Mine 0.0281871832982

Component Number:  2
6100 Snow Patrol: Chasing Cars 0.0379937354037
3238 Snow Patrol: Run 0.0351394834574
5976 Keane: Somewhere Only We Know 0.0346835391398
5697 Coldplay: Fix You 0.0340572087773
9269 The Killers: When You Were Young 0.03351177052
4690 The Fray: How To Save A Life 0.0329460616405
633 Coldplay: Viva La Vida 0.0321872938911
5694 Coldplay: Speed Of Sound 0.0314566624118
7185 Snow Patrol: Chocolate 0.0314232865266
```

7862 The Killers: Read My Mind 0.0311445037731
6102 Snow Patrol: You'Re All I Have 0.030293996609
9265 The Killers: Bones 0.0301016741267
2163 Mika: Grace Kelly 0.0300161921033
6093 Snow Patrol: Open Your Eyes 0.0294973318651
6245 Coldplay: Clocks 0.0293846185055
5120 Coldplay: The Scientist 0.0285674019869
5052 The Kooks: Naïve 0.0277590241249
8447 Franz Ferdinand: Take Me Out 0.0277189567374
28975 Maroon 5: She Will Be Loved 0.0269499599944
6088 Snow Patrol: How To Be Dead 0.0268015040182

Component Number: 3
6759 Boy Division: Love Will Tear Us Apart 0.0366855029336
18896 Pulp: Common People 0.0296103229186
6267 The Stone Roses: I Wanna Be Adored 0.0271893811421
10906 David Bowie: Heroes 0.0267410991145
24882 Joy Division: She'S Lost Control 0.0259614775707
3426 The Clash: London Calling 0.0258114333907
43810 New Order: Blue Monday 0.025639835559
43640 Joy Division: Transmission 0.0252786612713
10865 The Stone Roses: I Am The Resurrection 0.0243102627931
9446 The Smiths: This Charming Man 0.023973698677
24906 Pulp: Disco 2000 0.0238105642016
50749 David Bowie: Ashes To Ashes 0.0236049066902
45318 Blondie: Atomic 0.0235929495998
7725 Supergrass: Alright 0.0232342764614
44079 Roxy Music: Virginia Plain 0.0232174825066
10719 David Bowie: Ziggy Stardust 0.0231146947031
22345 Depeche Mode: Just Can'T Get Enough 0.0227791201851
10079 David Bowie: Changes 0.0227741624903
43249 The Undertones: Teenage Kicks 0.0226649451024
17791 Blondie: One Way Or Another 0.0225369925343

Component Number: 4
5369 Bloc Party: Like Eating Glass 0.0297049786374
5368 Bloc Party: Helicopter 0.0290584874318
6384 Death Cab For Cutie: Title And Registration 0.0258704962147
6377 Death Cab For Cutie: A Lack Of Color 0.0258264576919
5376 Bloc Party: This Modern Love 0.0253239112584
8595 Arctic Monkeys: I Bet You Look Good On The Dancefloor 0.0252503819043
11270 Interpol: Slow Hands 0.025072423086
5775 Bloc Party: Banquet 0.0239473332632
5367 Bloc Party: Positive Tension 0.0235461130635
4982 Death Cab For Cutie: I Will Follow You Into The Dark 0.0234082356027
3443 Muse: Supermassive Black Hole 0.0231070415051

```
3442 Muse: Map Of The Problematique 0.0229441672584
6378 Death Cab For Cutie: We Looked Like Giants 0.0225265142519
4980 Death Cab For Cutie: Soul Meets Body 0.0225161424396
6381 Death Cab For Cutie: Tiny Vessels 0.0225144350262
9056 The Strokes: Reptilia 0.0224666370983
5372 Bloc Party: So Here We Are 0.0224005450008
9595 Arcade Fire: Wake Up 0.0220345340829
9383 Arctic Monkeys: Mardy Bum 0.0218501057131
3749 Muse: Starlight 0.0218184582435

Component Number:  8
49424 Nine Inch Nails: Closer 0.0268224653831
56208 Nine Inch Nails: Every Day Is Exactly The Same 0.0197942440722
50317 The Smiths: How Soon Is Now? 0.0196700073077
18330 Interpol: Evil 0.0187394890809
56211 Nine Inch Nails: Only 0.0186701186915
21256 Nine Inch Nails: Hurt 0.0186171502139
56209 Nine Inch Nails: The Hand That Feeds 0.0184825533045
43313 Nine Inch Nails: Head Like A Hole 0.0175570227333
56206 Nine Inch Nails: Terrible Lie 0.0175138966689
57119 Nine Inch Nails: Sin 0.0174274680036
43312 Nine Inch Nails: March Of The Pigs 0.0172172316567
3443 Muse: Supermassive Black Hole 0.0169881386653
6761 The Smiths: Panic 0.0169172784127
55356 Nine Inch Nails: Wish 0.016717622841
5494 We Are Scientists: Nobody Move, Nobody Get Hurt 0.0166419794175
56231 Nine Inch Nails: Sunspots 0.0164653079272
48382 Nine Inch Nails: Piggy 0.0164468597002
56213 Nine Inch Nails: All The Love In The World 0.0163210443265
56230 Nine Inch Nails: The Line Begins To Blur 0.0163079622156
56232 Nine Inch Nails: Getting Smaller 0.0162796738468
```

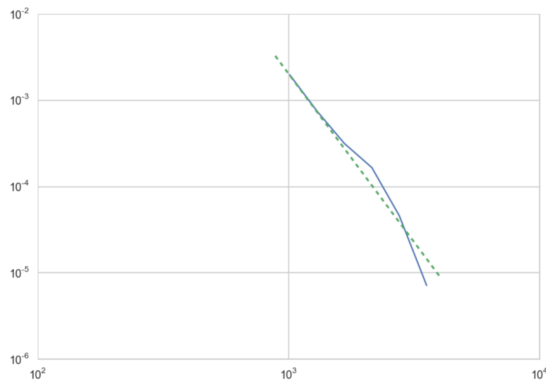
## k-Means

K-Means using the user play data as input yielded uninterpretable results.

## Distribution of Plays

Since we are trying to apply several NLP-based models to the data, one of the first steps is to compare the statistical properties of songs to words. Word frequencies are power law distributed, so a good start is to visualize the distribution of songs.

The following plots show the fit of the song and artist frequency distributions to a power law.

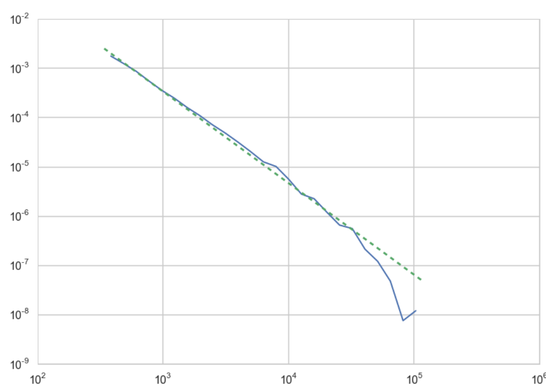


Power law exponent: 3.90072003565

Minimum x for power law fit: 885.0 out of 1498714

There is a bulge in the middle of the graph – perhaps this represents hipsters listening to relatively obscure songs quite a bit? Or, perhaps, a result of albums forcing a number of relatively obscure songs to be heard at once, making their frequencies higher than their rank would predict, given a power law distribution?

The tail end of the graph has a strong dropoff – many songs are listened to only once in the dataset. The exponent for song popularity is incredibly steep.

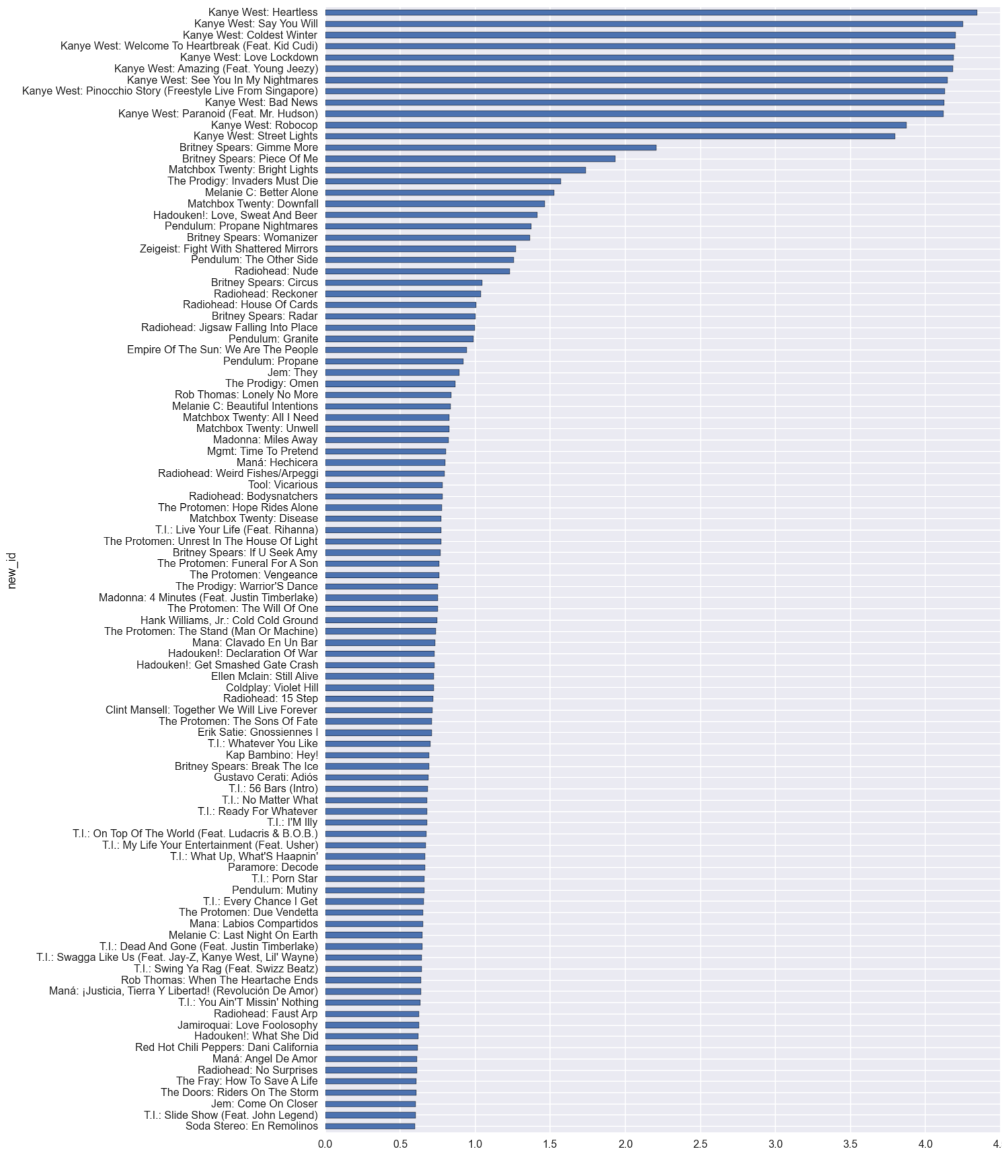


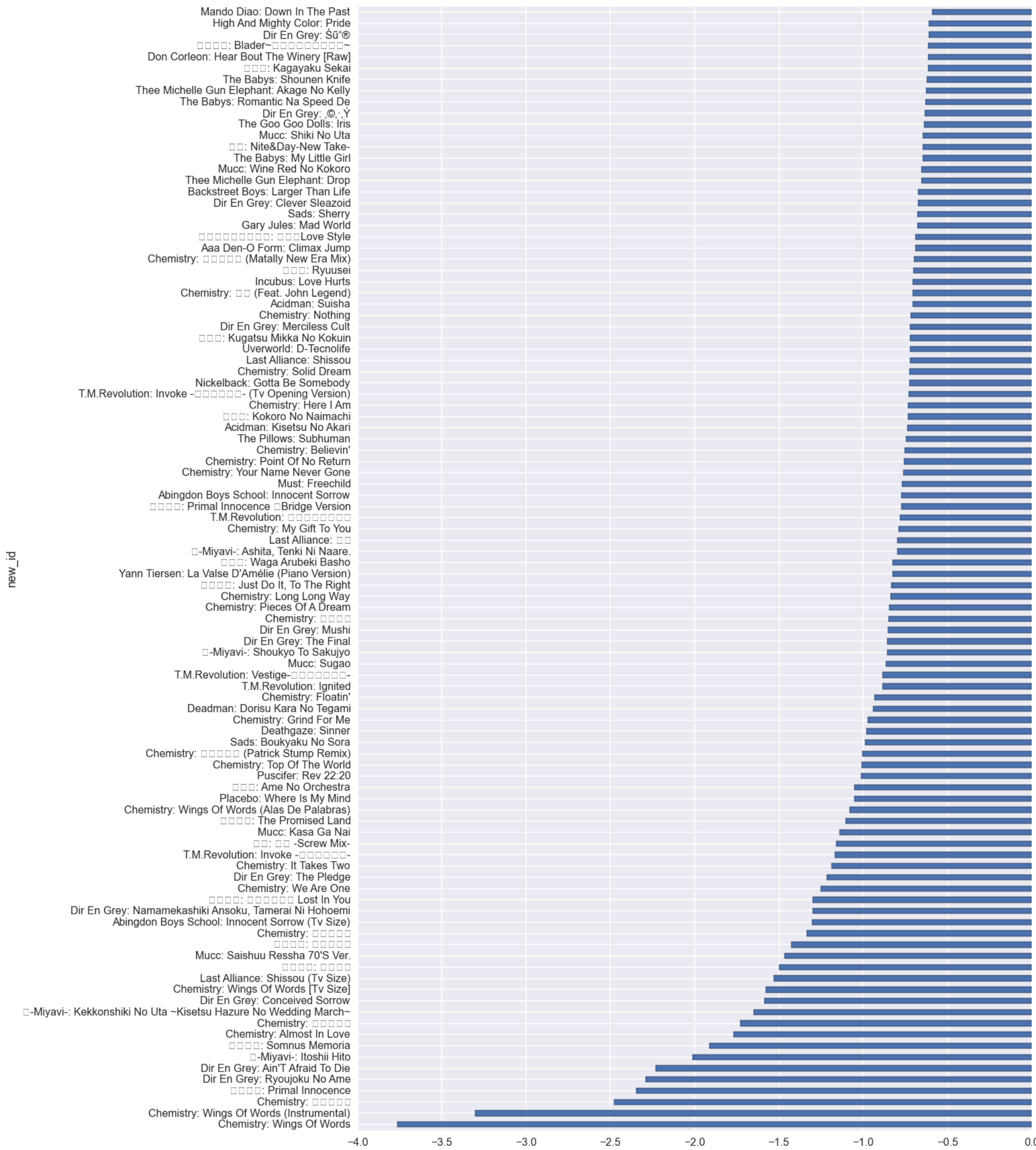
Power law exponent: 1.85803941411

Minimum x for power law fit: 340.0 out of 173919

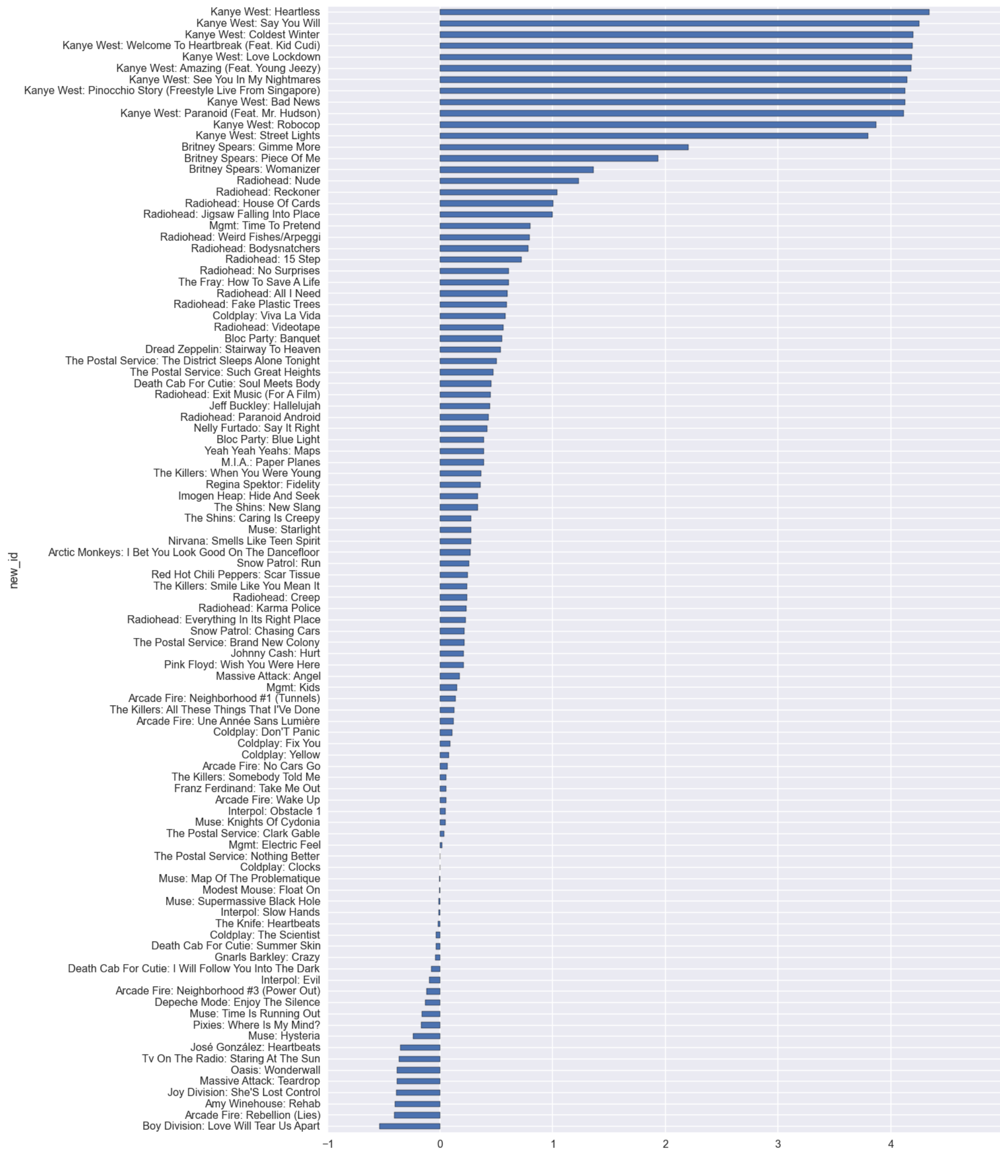
There is a strong dropoff at the bottom of the power law curve for artists.

The following bar charts [favored-by-males, favored-by-females] plot figures for the greatest difference between genders with respect to the question ?how many times was a song played by {gender} per person of {gender} in the dataset?" The other image, gender-skew plots the difference between the per capita gender statistics for the top songs in the dataset. However, since the dataset skews male, the ‘top songs’ are also skewed.


Kanye West: Heartless
Kanye West: Say You Will
Kanye West: Coldest Winter
Kanye West: Welcome To Heartbreak (Feat. Kid Cudi)
Kanye West: Love Lockdown
Kanye West: Amazing (Feat. Young Jeezy)
Kanye West: See You In My Nightmares
Kanye West: Pinocchio Story (Freestyle Live From Singapore)
Kanye West: Bad News
Kanye West: Paranoid (Feat. Mr. Hudson)
Kanye West: Robocop
Kanye West: Street Lights
Britney Spears: Gimme More
Britney Spears: Piece Of Me
Matchbox Twenty: Bright Lights
The Prodigy: Invaders Must Die
Melanie C: Better Alone
Matchbox Twenty: Downfall
Hadouken!: Love, Sweat And Beer
Pendulum: Propane Nightmares
Britney Spears: Womanizer
Zeigeist: Fight With Shattered Mirrors
Pendulum: The Other Side
Radiohead: Nude
Britney Spears: Circus
Radiohead: Reckoner
Radiohead: House Of Cards
Britney Spears: Radar
Radiohead: Jigsaw Falling Into Place
Pendulum: Granite
Empire Of The Sun: We Are The People
Pendulum: Propane
Jem: They
The Prodigy: Omen
Rob Thomas: Lonely No More
Melanie C: Beautiful Intentions
Matchbox Twenty: All I Need
Matchbox Twenty: Unwell
Madonna: Miles Away
Mgmt: Time To Pretend
Maná: Hechicera
Radiohead: Weird Fishes/Arpeggi
Tool: Vicarious
Radiohead: Bodysnatchers
The Protomen: Hope Rides Alone
Matchbox Twenty: Disease
T.I.: Live Your Life (Feat. Rihanna)
The Protomen: Unrest In The House Of Light
Britney Spears: If U Seek Amy
The Protomen: Funeral For A Son
The Protomen: Vengeance
The Prodigy: Warrior'S Dance
Madonna: 4 Minutes (Feat. Justin Timberlake)
The Protomen: The Will Of One
Hank Williams, Jr.: Cold Cold Ground
The Protomen: The Stand (Man Or Machine)
Mana: Clavado En Un Bar
Hadouken!: Declaration Of War
Hadouken!: Get Smashed Gate Crash
Ellen Mclain: Still Alive
Coldplay: Violet Hill
Radiohead: 15 Step
Clint Mansell: Together We Will Live Forever
The Protomen: The Sons Of Fate
Erik Satie: Gnossiennes I
T.I.: Whatever You Like
Kap Bambino: Hey!
Britney Spears: Break The Ice
Gustavo Cerati: Adiós
T.I.: 56 Bars (Intro)
T.I.: No Matter What
T.I.: Ready For Whatever
T.I.: I'M Illy
T.I.: On Top Of The World (Feat. Ludacris & B.O.B.)
T.I.: My Life Your Entertainment (Feat. Usher)
T.I.: What Up, What'S Haapnin'
Paramore: Decode
T.I.: Porn Star
Pendulum: Mutiny
T.I.: Every Chance I Get
The Protomen: Due Vendetta
Mana: Labios Compartidos
Melanie C: Last Night On Earth
T.I.: Dead And Gone (Feat. Justin Timberlake)
T.I.: Swagga Like Us (Feat. Jay-Z, Kanye West, Lil' Wayne)
T.I.: Swing Ya Rag (Feat. Swizz Beatz)
Rob Thomas: When The Heartache Ends
Maná: ¡Justicia, Tierra Y Libertad! (Revolución De Amor)
T.I.: You Ain'T Missin' Nothing
Radiohead: Faust Arp
Jamiroquai: Love Foolosophy
Hadouken!: What She Did
Red Hot Chili Peppers: Dani California
Maná: Angel De Amor
Radiohead: No Surprises
The Fray: How To Save A Life
The Doors: Riders On The Storm
Jem: Come On Closer
T.I.: Slide Show (Feat. John Legend)
Soda Stereo: En Remolinos
new_id
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5



new_id
Mando Diao: Down In The Past
High And Mighty Color: Pride
Don Corleon: Hear Bout The Winery [Raw]
The Babys: Shounen Knife
Thee Michelle Gun Elephant: Akage No Kelly
The Babys: Romantic Na Speed De
The Goo Goo Dolls: Iris
Mucc: Shiki No Uta
The Babys: My Little Girl
Mucc: Wine Red No Kokoro
Thee Michelle Gun Elephant: Drop
Backstreet Boys: Larger Than Life
Dir En Grey: Clever Sleazoid
Sads: Sherry
Gary Jules: Mad World
Aaa Den-O Form: Climax Jump
Incubus: Love Hurts
Acidman: Suisha
Chemistry: Nothing
Dir En Grey: Merciless Cult
Uverworld: D-Tecnolife
Last Alliance: Shissou
Chemistry: Solid Dream
Nickelback: Gotta Be Somebody
Chemistry: Here I Am
Acidman: Kisetsu No Akari
The Pillows: Subhuman
Chemistry: Believin'
Chemistry: Point Of No Return
Chemistry: Your Name Never Gone
Must: Freechild
Abingdon Boys School: Innocent Sorrow
Chemistry: My Gift To You
Yann Tiersen: La Valse D'Amélie (Piano Version)
Chemistry: Long Long Way
Chemistry: Pieces Of A Dream
Dir En Grey: Mushi
Dir En Grey: The Final
Mucc: Sugao
T.M.Revolution: Ignited
Chemistry: Floatin'
Deadman: Dorisu Kara No Tegami
Chemistry: Grind For Me
Deathgaze: Sinner
Sads: Boukyaku No Sora
Chemistry: Top Of The World
Puscifer: Rev 22:20
Placebo: Where Is My Mind
Chemistry: Wings Of Words (Alas De Palabras)
Mucc: Kasa Ga Nai
Chemistry: It Takes Two
Dir En Grey: The Pledge
Chemistry: We Are One
Dir En Grey: Namamekashiki Ansoku, Tamerai Ni Hohoemi
Abingdon Boys School: Innocent Sorrow (Tv Size)
Mucc: Saishuu Ressha 70'S Ver.
Last Alliance: Shissou (Tv Size)
Chemistry: Wings Of Words [Tv Size]
Dir En Grey: Conceived Sorrow
Chemistry: Almost In Love
Dir En Grey: Ain'T Afraid To Die
Dir En Grey: Ryoujoku No Ame
Chemistry: Wings Of Words (Instrumental)
Chemistry: Wings Of Words
−4.0
−3.5
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0




582111 distinct songs remained after thresholding for those listened to by both genders.

## Analysis of Topic Modeling

In a power law, the probability of a word being the xth word, where xs represent the vocabulary indexed by descending rank, is inversely proportional to its rank: $p(w = x) \propto x^{-l}$, where $l$ is an constant representing how steep the falloff is. In natural language, this is often referred to as Zipf's law.

Here are the results for topic modeling. They're kind of incoherent, but there's generally a coherent subset of songs within each topic (identified by playing random songs we don't know on Spotify). Unfortunately, it takes far too long to run to able to tweak it until it's perfect. This is actually really hard, since, with text, you can tell at a glance if topics are coherent.

In an attempt to identify user personalities, we split each user's listening history by month to create a collection of 22,858 documents. Songs were turned into numeric IDs, and these were used as tokens for the model.

Summary statistics for the number of songs played during each user-month:

```
mean 838.243460
std 1028.330952
min 1.000000
25% 172.000000
50% 521.000000
75% 1123.000000
max 12763.000000
```

Here are some sample topics.

```
Topic 1:
Rodgers And Hart: Nobody'S Heart - Doris Day
The Malleys: Coming Home
Rodgers And Hart: Bewitched - Frank Sinatra
Rodgers And Hart: Glad To Be Unhappy - Frank Sinatra
Rodgers And Hart: Mountain Greenery - Ella Fitzgerald
Rodgers And Hart: With A Song In My Heart - Perry Como
Pet Shop Boys: Absolutely Fabulous (7" Mix)
House Shoes: Midnight Running Club
Brass Construction: Got To Be Love
Kelly Senecal: Mvb Radio #9

Topic 2:
John Lennon: Hold On (Remix)
Mygermanclass.Com: Ubel Knubels Welt: Episode 1
Heartless Bastards: Pass And Fail
Frankie Knuckles: It'S A Cold World
Sugarcult: Stuck In America
The Ducky Boys: The War Back Home
Dj Mehdi: Love Bombing
The Beach Boys: When A Man Needs A Woman (Digitally Remastered 01)
Hypnosis Files: Trainmmo
The National: Looking For Astronauts

Topic 3:
Trans-X: Message On The Radio (Remix)
```

```
Isobel Campbell & Mark Lanegan: (Do You Wanna) Come Walk With Me?
Me First And The Gimme Gimmes: Summertime
Oasis: Supersonic
Paolo Fresu Quintet: Prayer For Sibylle
Richard Cheese: Holiday In Cambodia 2006
Adam Green: Broken Joystick
The Clash: Safe European Home
Nouvelle Vague: Heart Of Glass
Blonde Redhead: For The Damaged

Topic 4:
Sagor & Swing: Apollons Aftonsång
Supreme Beings Of Leisure: Truth From Fiction
Hezekiah: Hurry Up & Wait (Intro)
Frank Black: Rock A My Soul
Yo-Yo Ma & The Silk Road Ensemble: Chi Passa Per'Sta Strada (Filippo Azzaiolo)
Spiritualized: Angel Sigh
Mad Caddies: Depleted Salvo
Silversun Pickups: The Fuzz
Grandaddy: Everything Beautiful Is Far Away
Elvis Costello & The Attractions: Accidents Will Happen
```

**Why Results are Bad**

Hypothesis: Vanilla LDA, with its Dirichlet-Multinomial updating, is fine for language, but it does not capture the power-law tendencies of the data. Since the power law exponents we are dealing with here are much larger than those for language (which are barely above 1), the rich-get-richer phenomenon is more pronounced for songs.

Solution: try Pitman-Yor process topic models, which can be represented by a Chinese Restaurant Process. These can capture power-law data.

Other possibilities: better thresholding is needed to get rid of low-probability songs, which strongly identify documents, allowing LDA to find a good solution from a probabilistic point of view, but not for human understanding. Or, perhaps, a different splitting method is needed, if the documents are too similar or too different for LDA to capture useful information based on the statistical patterns of word co-occurrence within documents.

Shalit et al. 2013: Modeling Musical Influence with Topic Models. This article performs topic modeling on extracted audio features for songs in order to show the evolution of popular music over time, but similar methods could be used to model the evolution of individual users over the dataset.

They also model the influence of songs on other songs, where the topic-word distribution vectors move according to a random walk at every timestep, where the variance is a topic evolution speed parameter, and the mean of each new topic is based on a combination of the previous topic distribution and the other topics weighted by influence scores.

Several evaluation metrics have been designed for topic modeling, although many require expert annotations of topics. Mimno et al. 2011: Optimizing Semantic Coherence in Topic Models suggests that traditional approaches to evaluating ML models do not produce good topics (e.g. predicting held-out documents). Instead, evaluation methods based on co-occurrence statistics were shown to have a high correlation with human judgements of topic "coherence."

Based on this article, and previous research, we plan on using the following evaluation metric for topic models:

The average pairwise PMI for songs within a topic, where PMI, or pointwise mutual information is defined as follows:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Intuitively, a high average topic PMI means that songs in a topic are much more likely to co-occur in a user's listening than their frequencies alone would suggest.

PMI can be calculated from the dataset itself, where co-occurrence is determined based on some as-of-yet unknown method (Maybe, songs listened to by a user within the same day? 15-song sliding window? Should be significantly smaller than the document level. Co-occurrence can, and perhaps should, be determined by external data.)

One huge problem I foresee: LDA assumes a bag-of-words model, so order within a document is meaningless. Additionally, the graphical model (https://mollermara.com/blog/lda/lda-tikz.png)for LDA ASSUMES CONDITIONAL INDEPENDENCE OF TOKENS GIVEN TOPICS AND THE TOPIC PROPORTIONS OF A DOCUMENT.

This assumption is violated by albums: a set of (usually similar) songs that (usually) occur together. LDA can tend to put the songs in an album in a topic, since that's a good solution to the optimization problem.

Solution: there are versions of LDA that can detect 'burstiness'. There are also versions of LDA that can detect phrases – where a "phrase" is a bunch of songs that occur in a row, which can be an album or a subset thereof.

# 4 Modeling the Data

It seems natural to model our data in the topic models framework. There are 'topics' of song, hopefully corresponding to genre. Users will be the document, and they will have some topic distribution. If 1000 documents isn't enough, we'll split users into multiple documents in some reasonable way. In vanilla topic modeling, words (here songs) are drawn independent of the previous word, but playlists are incredibly sticky. That is, if the last song you listened to was by Kanye, the next song will be by Kanye too, or something really close. Thus, we may want to think of each session start as a draw from the user's distribution of topics. This can be modeled by adding very heavy sticky Markovian dependency on the previously played song.

The next step would be to allow the users' topic distribution to change over time, and allow the topics themselves to change too. We have time series data, so as new songs appear, they may create new genres or change existing ones. Allowing this freedom may lead to interesting observations, such as perhaps finding songs that change the landscape of popular music.

# 5 Some Prior Works

The following paper considers the evolution of popular music from 1960-2010, using topic modeling.
`http://rsos.royalsocietypublishing.org/content/royopensci/2/5/150081.full.pdf`

This paper also used topic modeling to cluster songs.
`http://jmlr.org/proceedings/papers/v28/shalit13.pdf`

Here is a paper using topic modeling on audio video data, heavily leveraging the time component of the data.
`http://www.idiap.ch/ odobez/publications/VaradarajanEmonetOdobez-IJCV-2013.pdf`

This paper studies how user opinions change as time goes on and they 'mature' in their community. Unsurprisingly, users tend to become more like their community or cluster the longer they stay there. While the ideas in this paper may not be directly useful if we don't work with text reviews, this idea of becoming more like the average member of your cluster is fascinating.
`http://cs.stanford.edu/people/jure/pubs/beerrec-www13.pdf`