Statistical Analysis of Network Data

Deezer : Comparaison de graphes et système de recommandation

Zakarya Ali
Vincent Le Meur

18 mai 2018

Table des matières

Introduction

Nous avons choisi d'étudier plusieurs réseaux provenant des données de l'application de streaming de musique Deezer. En plus de permettre aux utilisateurs d'écouter des chansons, l'application est aussi un réseau social où les gens peuvent se lier d'amitié. Les données que nous analysons sont issues du "Stanford Network Analysis Project (SNAP)" et ont été utilisé dans le cadre de l'article de Benedek Rozemberczki, Ryan Davies, Rik Sarkar et Charles Sutton : GEMSEC: Graph Embedding with Self Clustering.

Les données se décomposent en trois datasets, correspondant à trois pays différents : la Roumanie, la Croatie et la Hongrie. Les données représentent d'une part les relations d'amitié entre utilisateurs du réseau mais également leurs goûts musicaux classés par ordre de préférence.

Ainsi, l'objet de notre étude sera de comparer ces trois datasets en terme d'influence des goûts musicaux sur sa structure et d'utiliser cette compréhension du réseau pour mettre en place un système de recommandation. Pour cela nous représenterons les données des différents pays et nous les comparerons en utilisant plusieurs métriques.

1 Etude des datasets

Nous avons récupéré les données du tableau 1. Le premier constat est que le nombre de nodes et d'edges varient énormément d'un pays à un autre. En effet, la Croatie a un nombre d'edges et de nodes bien supérieur à ceux de la Roumanie et de la Hongrie.

Pays	Nodes (Utilisateurs)	Edges (Amitié)
Roumanie	41 773	125 826
Croatie	54 573 (+**31%**)	498 202 (+**296%**)
Hongrie	47 538 (+**14%**)	222 887 (+**77%**)

TABLE 1: Nodes et edges par pays

Pour étudier en profondeur ces données, notre première idée était de représenter les données dans leur ensemble. Par exemple, nous avons tenté de représenter chacun des utilisateurs avec leur liens d'amitié et une couleur associée à leur style préféré (voir figure 1). Le code que nous avons utilisé pour établir l'ensemble de notre étude est disponible ici : `https://github.com/zakaryaxali/network_deezer`.
Le manque de visibilité et d'interprétation claire de ce type de graphique nous a donc poussé à réfléchir à un moyen de simplifier la représentation des données. Nous en avons étudié deux : une représentation par "superusers" et un regroupement des données par genres musicaux.

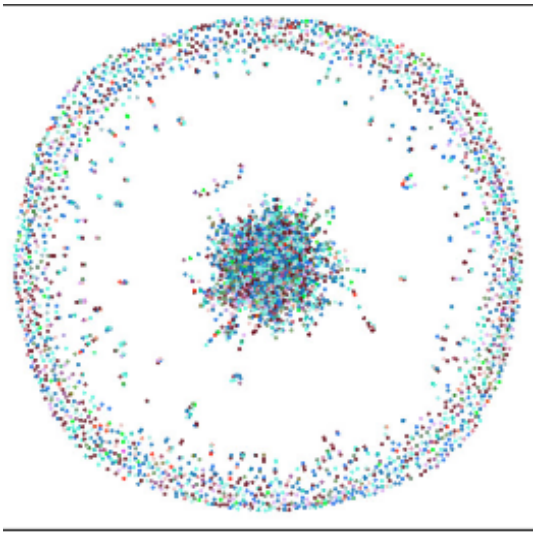


FIGURE 1: Représentation des liens d'amitié de 5000 utilisateurs de Deezer en Croatie

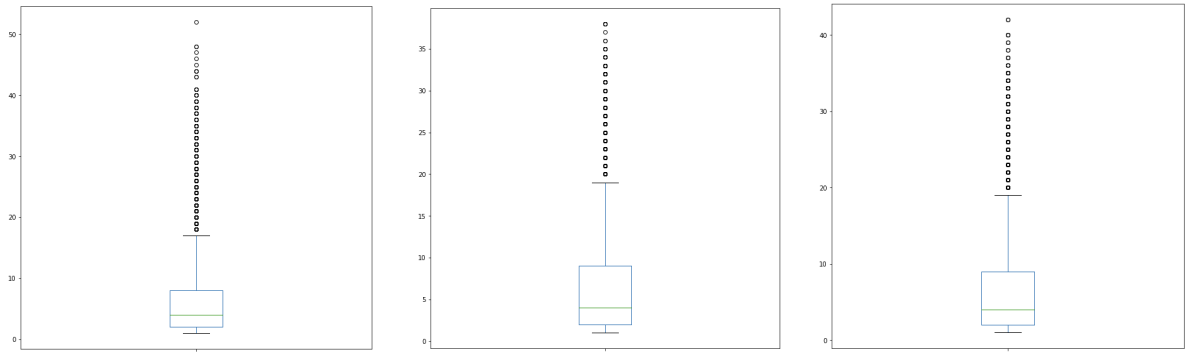
1.1 Super users

Devant la quantité d'informations fournies sur les utilisateurs, nous nous sommes demandées dans quelle mesure nous pouvions en garder autant que possible dans la représentation des réseaux. Après avoir étudié le nombre moyen de genres favoris pour chaque utilisateur (voir figure 2 et table 2) nous avons choisi d'utiliser des super users pour réprésenter les liens. Nous définissons un super user comme un ensemble d'utilisateurs ayant leur 3 premiers styles en commun.

Croatie	Roumanie	Hongrie
6.3	6.04	5.99

TABLE 2: Nombre moyen de genres aimés des utilisateurs, par pays

L'idée est d'analyser la dynamique que peut apporter le fait d'avoir plusieurs genres en commun. En regroupant les utilisateurs de cette façon on constate des rassemblements très hétérogènes. On a essayé de montrer les liens entre super users dans chaque pays avec une représentation aléatoire. La figure 3 montre cette représentation pour les 3 pays où la taille d'un noeud (super user) est proportionnelle au nombre d'utilisateurs qu'il contient et l'épaisseur des edges est proportionnel au nombre d'amis qu'ont en commun chaque super users liés. Malheureusement, cette représentation est très difficile à lire puisqu'il y a beaucoup de super users. Il faut une représentation plus concise pour représenter nos données. (Trop volumineuse pour être dans le rapport, on peut retrouver les représentations détaillées des super users pour chaque pays ici : la Roumanie, la Croatie et la Hongrie)



(a) Roumanie (b) Croatie (c) Hongrie

FIGURE 2: Boxplot nombre de genres aimés par utilisateurs

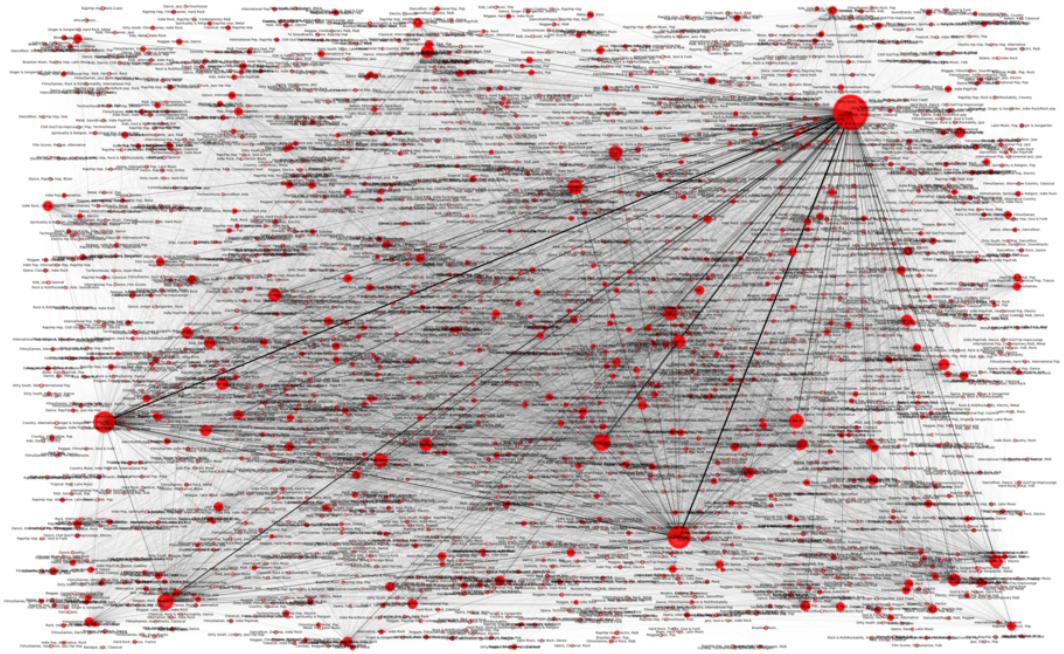


FIGURE 3: Représentation des super users (Roumanie)

1.2 Représentation par genres

Pour la représentation par genres, on souhaite placer chaque utilisateur dans un cluster associé à son style de prédilection. Nous voulons également que les liens entre chacun de ces clusters prennent en compte les relations d'amitié du réseau entre utilisateurs. Nous avons amis en place une matrice empirique $Mprob$ de "probabiltié d'amitié entre styles".

Chaque ligne et chaque colonne de la matrice représente un style musical du dataset. Ainsi $Mprob_{[j,k]}$ représente une probabilité empirique d'amitié avec le cluster k sachant qu'on appartient au cluster j. Elle se calcule à partir de l'ensemble des relations d'amitié de chaque cluster.

L'idée est alors d'observer la répartition des liens d'amitié de tous les utilisateurs de chaque clusters avec les autres clusters. Tout d'abord, nous avons mis en place une matrice "sparse" M_{sparse} de dimension $n_{users} \times n_{users}$ telle que $M_{sparse}[j,k] = 1$ si j et k sont amis et 0 sinon. Formellement, nous avons ensuite calculé $Mprob$ de la manière suivante :

$$Mprob_{[i,k]} = \sum_{j \in S(i)} \left(\frac{\sum_{l \in S(k)} M_{sparse}[j,l]}{\sum_{k=1}^{n_{styles}} M_{sparse}[j,k]}\right)$$

avec $S(i)$ et $S(k)$ l'ensemble des utilisateurs respectivement associés au cluster i et k, et n_{styles} le nombre de styles différents du dataset.

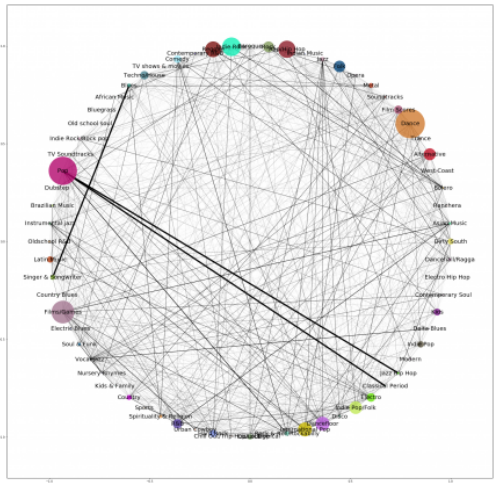
Nous avons enfin normalisé (en divisant par la somme de chaque ligne) les lignes de $Mprob$ pour obtenir des probabilités comprises entre 0 et 1.

Cette matrice va être utile pour représenter les liens entre clusters de manière pondérée ainsi que pour la recommandation.

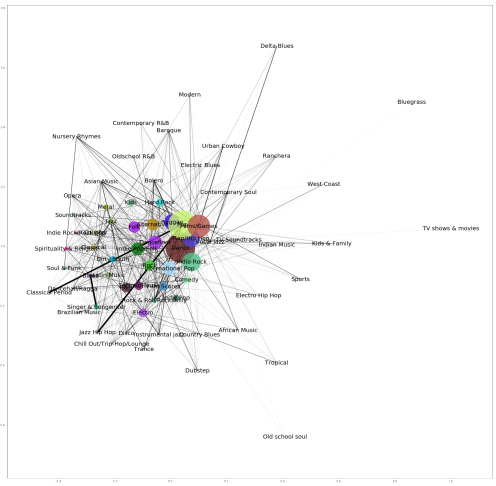
Nous allons représenter les graphes obtenus pour chaque pays. Un noeud correspond à un style musical et sa taille est proportionnelle au nombre de users. Les liens sont obtenus à partir de la matrice $Mprob$: plus le lien entre deux styles i et k est important plus la valeur correspondante $Mprob[i,k]$ est élevée et plus le lien est épais. Nous utilisons deux façons de représenter le graphe :

- Une représentation circulaire de chacun des noeuds avec les liens en position centrale
- Une représentation utilisant l'algorithme : "Fruchterman-Reingold force-directed algorithm".

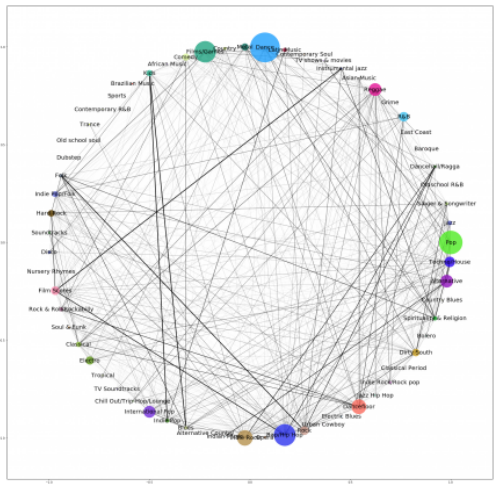
Nous avons ainsi, pour chaque pays, obtenu les graphes en figure 4.
On constate que le second type de représentation (de Fruchterman-Reingold) a tendance à rassembler au centre les styles ayant des liens importants et à excentrer les styles plus isolés. Bien qu'intéressant pour l'exemple de la Croatie, cette représentation n'est pas lisible pour les deux autres pays. En revanche, la représentation en cercle nous permet de constater que les différents clusters semblent plus ou moins de même taille entre différents pays mais comportent des liens différents d'un pays à l'autre : on constate par exemple un lien très



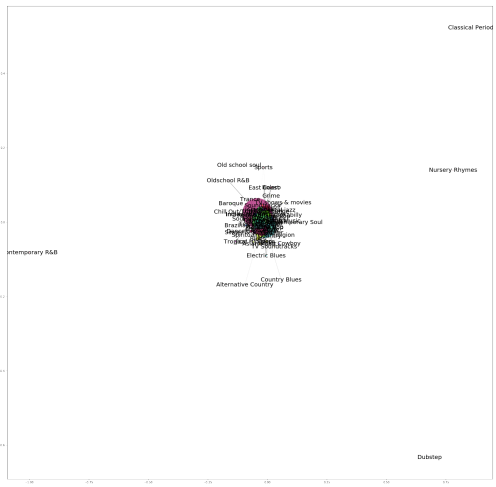
(a) Circulaire (Croatie)



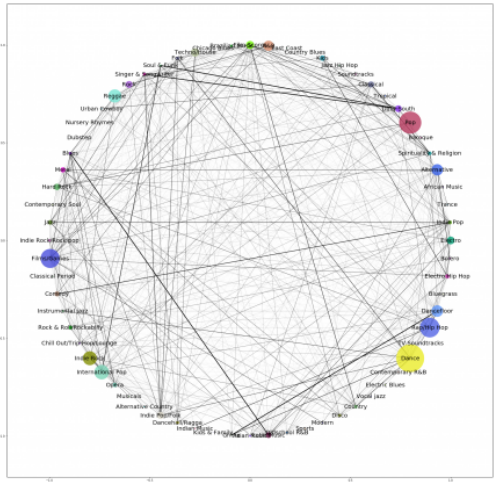
(b) Fruchterman-Reingold (Croatie)



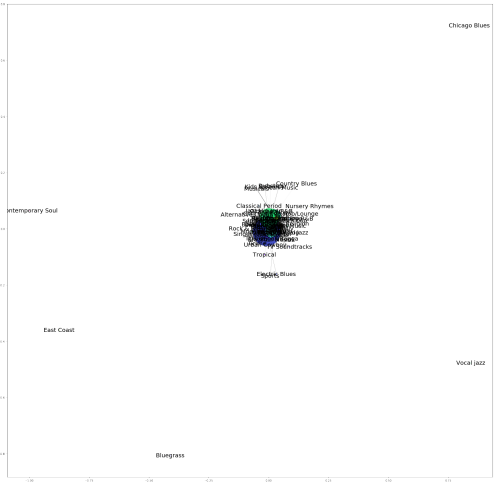
(c) Circulaire (Hongrie)



(d) Fruchterman-Reingold (Hongrie)



(e) Circulaire (Roumanie)



(f) Fruchterman-Reingold (Roumanie)

FIGURE 4: Différentes représentations des datasets selon les genres

fort entre la Pop et le Jazz/Hip Hop pour le dataset croate alors que ce lien est bien moins important pour les autres datasets.

2 Comparaison des datasets

L'objet de cette partie est d'effectuer une comparaison entre les différents datasets. Nous allons les comparer à partir des deux modélisations détaillées dans la partie précédente. Ainsi la première comparaison se basera sur des statistiques issus de la modélisation "super-users" alors que la seconde portera sur une étude de la distribution en degré d'une nouvelle modélisation en "superstyles".

2.1 Super users

On présente dans la table 3 les 5 plus importants super users (en terme de nombre d'utilisateurs qu'ils contiennent) par pays. Pour tous les pays le super user dominant est la Pop. En effet, plus de 10% des utilisateurs de chaque pays n'ont que ce genre de musique favoris. Ensuite, on observe que la Roumanie et la Hongrie ont, dans le même ordre, les 3 même premiers super users. "Pop", "Rap/Hip-Hop" et "Dance, Rap/Hip-Hop, Pop". La Croatie favorise davantage le Rock. De cette comparaison, il émerge qu'il est difficile d'établir un lien clair entre les goûts de la Croatie et des 2 autres pays. De plus, on constate qu'un grand nombre d'utilisateurs dans ces pays n'ont pas plus d'un genre préféré. La piste du super user est donc difficile à interpréter.

Roumanie	
Genre(s)	Utilisateurs
Pop	4882
Rap/Hip-Hop	1844
Dance, Rap/Hip-Hop, Pop	1620
Dance, Pop	950
Dance	831

Hongrie	
Genre(s)	Utilisateurs
Pop	5498
Rap/Hip-Hop	2013
Dance, Rap/Hip-Hop, Pop	1877
Dance	1224
Rap/Hip-Hop, Pop	1098

Croatie	
Genre(s)	Utilisateurs
Pop	6868
Dance, Rap/Hip-Hop, Pop	1674
Pop, Rock	1147
Rock	1014
Indie Rock, Indie Pop/Folk, Dance	937

TABLE 3: Précision des algorithmes pour la classification binaire

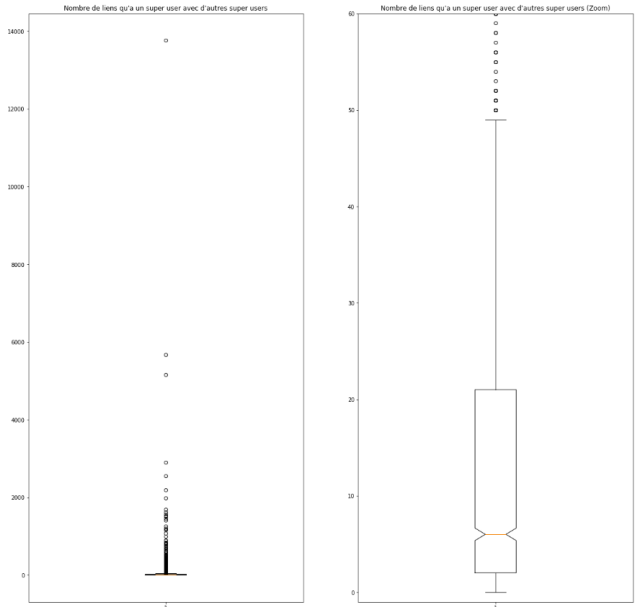
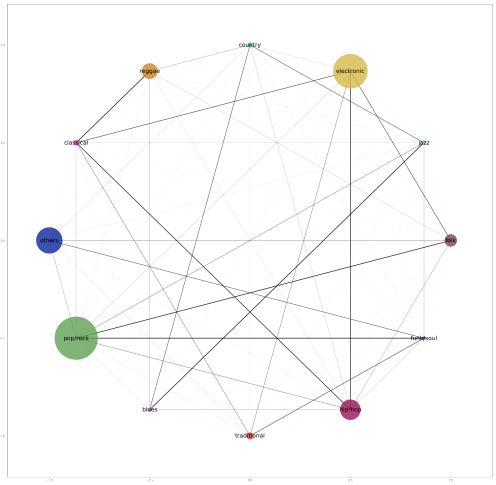


FIGURE 5: Représentation des super users (Roumanie)

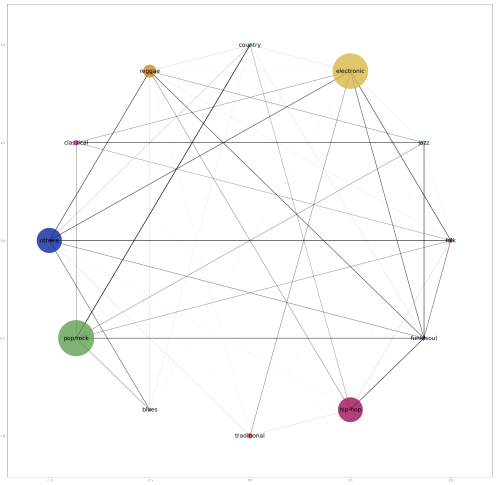
2.2 Genres musicaux

La modélisation par genre effectuée dans la section I.2 ne nous permet pas de comparer efficacement les différents pays : le graphe simplifié est encore trop complexe pour une analyse détaillée et certains styles sont présents seulement dans un pays ce qui gêne toute comparaison. Nous avons donc eu l'idée de créer des "superstyles". Nous avons plus ou moins arbitrairement regroupé les dizaines de styles existants en 12 "superstyles" : folk, jazz, electronic, country, reggae, classical, others, pop/rock, blues, traditional, hip-hop, et funk/soul. Nous avons ensuite mené la même étude que précedemment et obtenu les graphes suivants (pour la représentation circulaire) :
Ici encore la taille des clusters varie peu d'un pays à l'autre mais ce sont les liens entre clusters qui varient : nous avons par exemple un lien très fort entre reggae et funk/soul pour la Roumanie qu'on ne retrouve pas dans d'autres pays.

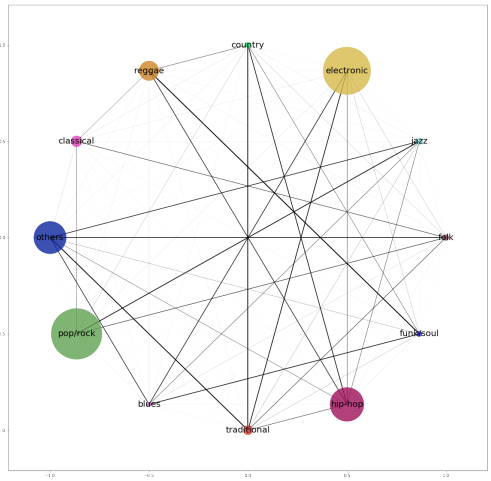
Nous allons à présent analyser la distribution en degrés dans chaque superstyle. Concrètement nous allons pour chaque cluster évaluer le degré de chaque utilisateur et représenter cette distribution en degré par un boxplot par superstyle. Nous obtenons alors :
Nous constatons alors que la Croatie a systématiquement une distribution plus étendue que les deux autres pays avec une médiane plus élevée pour l'ensemble des superstyles. La Roumanie et la Hongrie ont une distribution plutôt proche en étendue et en médiane. Cette hétérogénéité entre datasets nous pousse alors pour notre dernière partie à effectuer une recommandation par pays (plutôt qu'une recommandation commune aux trois pays).



(a) Croatie



(b) Hongrie



(c) Roumanie

FIGURE 6: Représentations circulaires selon les superstyles

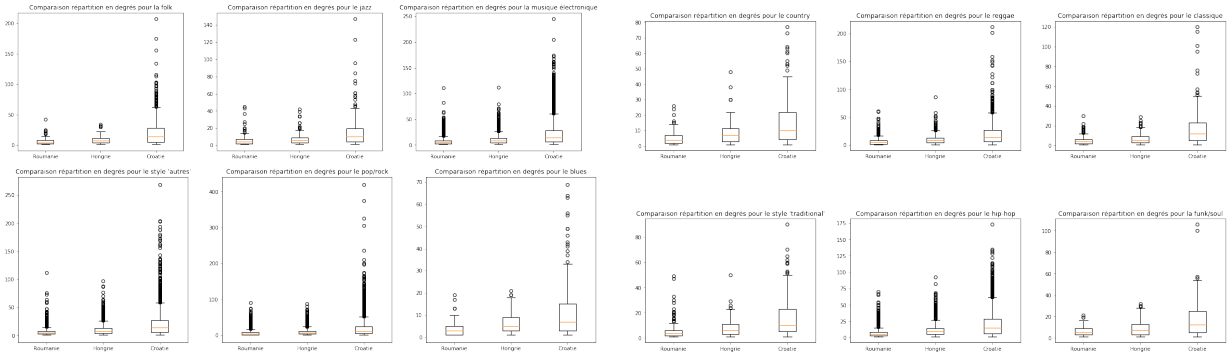


FIGURE 7: Comparaison des degrés des superstyles

3 Système de recommandation

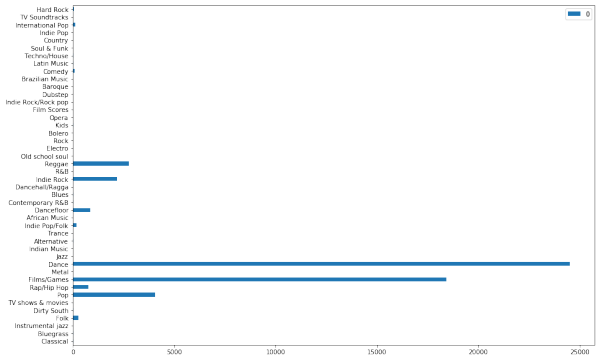
Dans cette partie nous allons mettre en place une recommandation pour nos utilisateurs. Nous avons décidé de traiter exclusivement une recommandation de nouveau goût musical.

Cette recommandation va se baser essentiellement sur la matrice $Mprob$ calculée sur chaque pays et sur les goûts de l'utilisateur. Cette utilisation de $Mprob$ nous semble intéressante car elle permet de définir une similarité entre goûts musicaux basés uniquement sur les liens sociaux des utilisateurs.

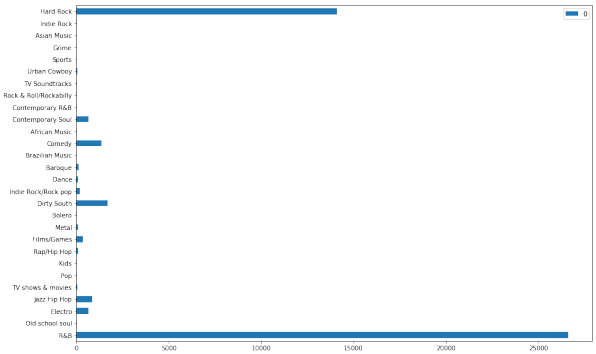
Le système de recommandation va fonctionner de la manière suivante :

- Sélectionner le style préféré de l'utilisateur
- Observer dans la matrice $Mprob$ le style le plus "proche" (probabilité la plus élevée de relation)
- Si l'utilisateur n'écoute pas ce genre musical on lui recommande sinon on passe au style préféré suivant de l'utilisateur
- Si tous les styles les plus proches sont déjà écoutés par l'utilisateur, on passe au deuxième style le plus proche (et ainsi de suite)

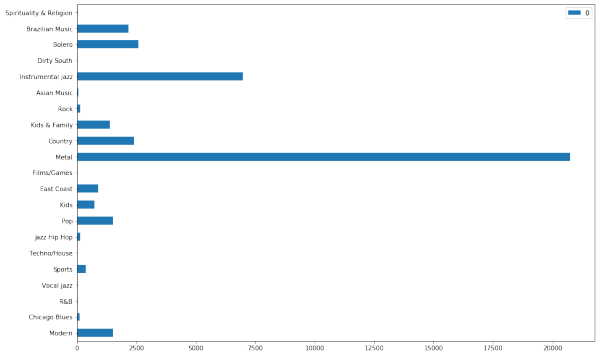
Nous avons alors lancer cette recommandation pour l'ensemble des utilisateurs des trois pays. Nous représentons ci-dessous l'histogramme de chaque pays des recommandations faites à chaque utilisateur :



(a) Croatie



(b) Hongrie



(c) Roumanie

FIGURE 8: Histogramme des recommandations

On constate d'abord que les recommandations sont différentes pour chaque pays : cela illustre encore une fois (a posteriori) le choix d'effectuer un système de recommandation par pays. Il y également à chaque fois deux styles majoritairement recommandés ("Dance" et "Films/Games" pour la Croatie, "R&B" et "Hard Rock" pour la Hongrie et enfin "Metal" et "Instrumental Jazz" pour la Roumanie).

Conclusion

En étudiant les réseaux d'utilisateurs, nous avons pu constater une hétérogénéité des genres musicaux majoritaires et les liens d'amitié en fonction de ces derniers pour les 3 datasets observés. Notre idée originale d'établir un système de recommandations global à tous les pays s'est vu contrarié par cette analyse et nous avons donc opté pour un système spécifique à chaque pays. Les résultats obtenues sont intéressants et peuvent être améliorés par exemple, en ayant des données de ce que les utilisateurs n'aiment pas. En effet, il serait alors possible de mettre en place un filtrage collaboratif, qui permettrait de mettre en jeu de façon plus direct les liens d'amités entre les utilisateurs pour apporter une recommandation moins standardisée.

Un autre aspect que nous avons souhaité explorer est à l'inverse, la recommandation d'ami, ou comment l'influence des goûts musicaux et des amis existants permet de proposer de nouvelles amitiés.

Une autre piste est l'analyse des genres entre eux, l'analyse d'une éventuelle hiérarchie au travers des genres aimés par les utilisateurs.

Annexe

Genres
Pop
Indie Rock
Indie Pop/Folk
International Pop
Rap/Hip Hop
Rock
Indie Pop
Alternative
Dance
Jazz
Techno/House
Electro
Singer & Songwriter
Films/Games
Contemporary R&B
R&B
Film Scores
Reggae
Folk
Disco
Contemporary Soul
Tropical
Brazilian Music
Dub
Asian Music
Electric Blues
Dancehall/Ragga
Electro Hip Hop

Genres
Musicals
Metal
Vocal jazz
Latin Music
Old school soul
Dirty South
Dancefloor
Hard Rock
Kids
Rock & Roll/Rockabilly
Blues
Comedy
Dubstep
Electro Pop/Electro Rock
Jazz Hip Hop
Soul & Funk
Classical
Country
Trance
Indie Rock/Rock pop
East Coast
Soundtracks
Chill Out/Trip-Hop/Lounge
TV Soundtracks
Spirituality & Religion
Opera
Ska

Genres
Bluegrass
Sports
Instrumental jazz
West Coast
Chicago Blues
Oldschool R&B
Country Blues
African Music
Acoustic Blues
Alternative Country
Game Scores
Bolero
Nursery Rhymes
Baroque
Urban Cowboy
Grime
Traditional Country
Bollywood
Ranchera
Old School
Indian Music
Early Music
Classical Period
Delta Blues
Modern
TV shows & movies
Classic Blues
Romantic
Kids & Family

TABLE 4: Liste des genres musicaux

Classical
Classical
Opera
Baroque
Early Music
Classical Period

Folk
Singer & Songwriter
Folk

Hip-Hop
Rap/Hip Hop
Contemporary R&B
R&B
Dirty South
Jazz Hip Hop
East Coast
West Coast
Oldschool R&B
Grime

Traditional
Latin Music
Tropical
Brazilian Music
Asian Music
African Music
Bolero
Nursery Rhymes
Bollywood
Ranchera
Indian Music

Blues
Blues
Electric Blues
Chicago Blues
Country Blues
Acoustic Blues
Delta Blues
Classic Blues

Funk/Soul
Disco
Old school soul
Soul & Funk
Contemporary Soul

Others
Films/Games
Film Scores
Kids
Comedy
Soundtracks
TV Soundtracks
Musicals
Spirituality & Religion
Sports
Game Scores
TV shows & movies
Romantic
Kids & Family

Electro
Dance
Techno/House
Electro
Dancefloor
Dubstep
Electro Pop/Electro Rock
Trance
Chill Out/Trip-Hop/Lounge
Dub
Electro Hip Hop

Country
Country
Bluegrass
Alternative Country
Urban Cowboy
Traditional Country

Jazz
Jazz
Vocal jazz
Instrumental jazz

Pop/Rock
Pop
Indie Rock
Indie Pop/Folk
International Pop
Rock
Indie Pop
Alternative
Metal
Hard Rock
Rock & Roll/Rockabilly
Indie Rock/Rock pop
Old School
Modern

Reggae
Reggae
Dancehall/Ragga
Ska

TABLE 5: Liste des genres hiérarchisés