Recommandation algorithmique et diversité de l'information

Comment analyser l'impact des algorithmes en ligne ?

Fabien Tarissan

CNRS – ENS Paris-Saclay

Internet & le web

Internet et le web

Internet

1966	projet Arpanet
29 oct. 1969	premier message

↦ *réseau de 4 sommets*

1971	projet Cyclade
1973	protocole TCP/IP (Vint Cerf & Robert Kahn)

↦ *réseaux de ∼ 50 de sommets*

1er jan. 1983	Arpanet adopte TCP/IP

↦ *réseaux de ∼ 1000 de sommets*

1987	Internet dépasse les 20 000 routeurs.

Le web

mars 1989	projet *World Wide Web* du CERN (Tim Berners-Lee & Robert Cailliau)
20 déc. 1990	1ère page web
30 avril 1993	CERN renonce aux droits

↦ *600 pages webs*

1994	*Netscape* et *Yahoo!*
1998	*Google*
2000 – ...	*Facebook*, *Youtube*, *Twitter*, ...
2007 – ...	*Deezer*, *Spotify*, *Netflix*, ...

En quête de visibilité

Tous les deux jours, nous créons autant d'information que l'humanité tout entière entre l'aube de son histoire et l'année 2003.

Eric Schmidt (PDG Google) en 2010

Le web en chiffres (dec. 2022) `http://www.internetlivestats.com/`

- 5,5 milliards d'utilisateur
- 1,9 milliards de sites web
- 570 millions de tweets et 5 milliards de vidéos par jours
- 5,6 milliards de requêtes et 200 milliards de mails envoyés par jours

$\Longrightarrow$ Nécessité d'algorithmes de classement

Quel impact sur l'information rendue visible ?

Quelle évolution ?

Dominique Cardon, À quoi rêvent les algorithmes. Nos vies à l'heure des big data, Paris, Seuil, La République des idées, 2015, 105 p.

Évolution des algorithmes classements de l'information :

Popularité (À côté du web) : nombre de réferences.
Clics, views, ..

Autorité (Au dessus du web) : authorité des réferences
Links $\Longrightarrow$ PageRank

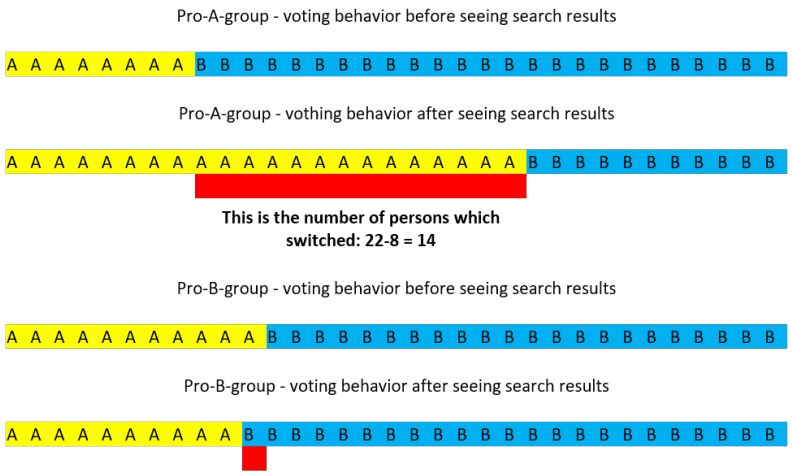
Notoriété (À l'intérieur du web): affinité (et fraicheur !) avec l'information issue de son voisinage.
Tweet, likes, ... $\Longrightarrow$ EdgeRank

Prédiction (Sous le web) : information inférée par le comportement d'un utilisateur.
Traces $\Longrightarrow$ Youtube / Spotify / Deezer / Netflix / ...

Quel impact ont ces algorithmes

Search Engine Manipulation effect

The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Robert Epstein and Ronald E. Robertson. PNAS 112 (33), 2015 (doi:10.1073/pnas.1419828112).



Résultats

- Classements biaisés ont un impact sur les votants (indécis) : $\geq 25\%$
- Les moteurs de recherche biaisés sont indétectés

À mettre en perspective avec :

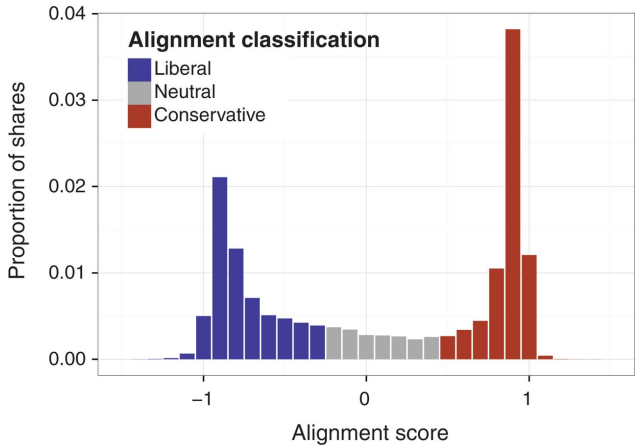
- 95% de la navigation web se fait sur 0.03% des pages existantes ...
- **The Fake News Machine. How Propagandists Abuse the Internet and Manipulate the Public**, Gu, Kropotov and Yarochkin, Trendlabs research paper, Trend Micro, 2015.

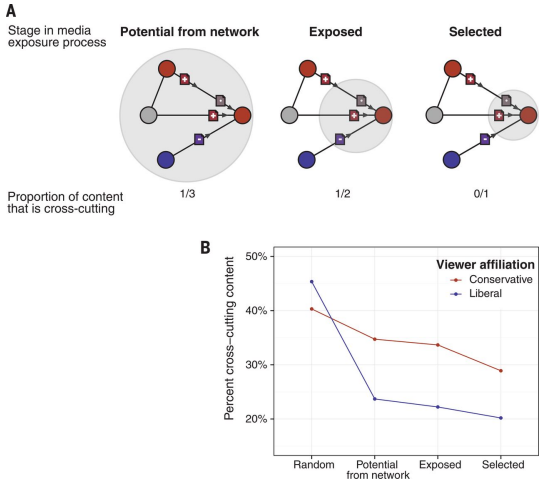
Chambres d'écho & bulles filtrantes

A squirrel dying in front of your house may be more relevant to your interests right now than people dying in Africa.

M. Zuckerberg.

Exposure to ideologically diverse news and opinion on Facebook. Eytan Bakshy, Solomon Messing, Lada A. Adamic. Science, 348 (6239), 2015 (doi:10.1126/science.aaa1160).





Résultats (N = 10.1 M)

- Les chambres d'échos sont dues principalement aux relations sociales
- Petit effet du filtrage algorithmique

école normale supérieure paris–saclay

Graphes & diversité

Mesurer la diversité

But : exploiter les structures relationnelles biparties induites par les interactions utilsateurs/plateformes.

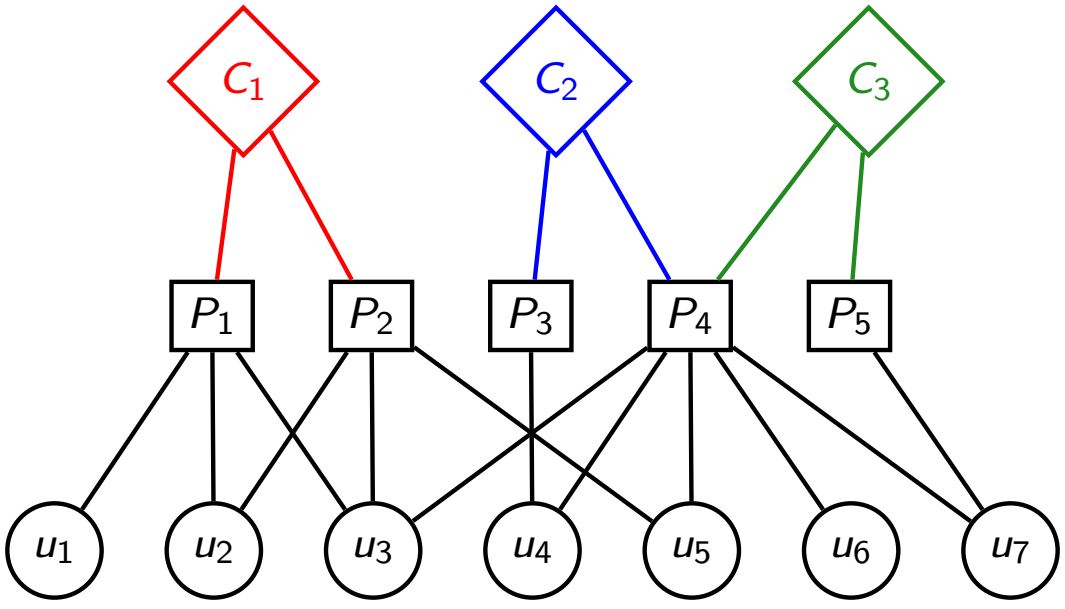
Systèmes de recommandations :

- Recommandations économiques (Amazon, ...)
- Recommandations informationnelles (Newsfeed de FB, média traditionnels, ...)

Représentation des données :

- Graphes bipartis : $\mathbb{B} = (\top, \bot, E_{\mathbb{B}})$
  - $\top$: produits (ou articles)
  - $\bot$: utilisateurs (ou lecteurs)
  - $E_{\mathbb{B}} \subset \top \times \bot$: relations produits/utilisateurs (ou articles/lecteurs)
- 2 graphes bipartis liés :
  - utilisateurs/produits
  - produits/catégories

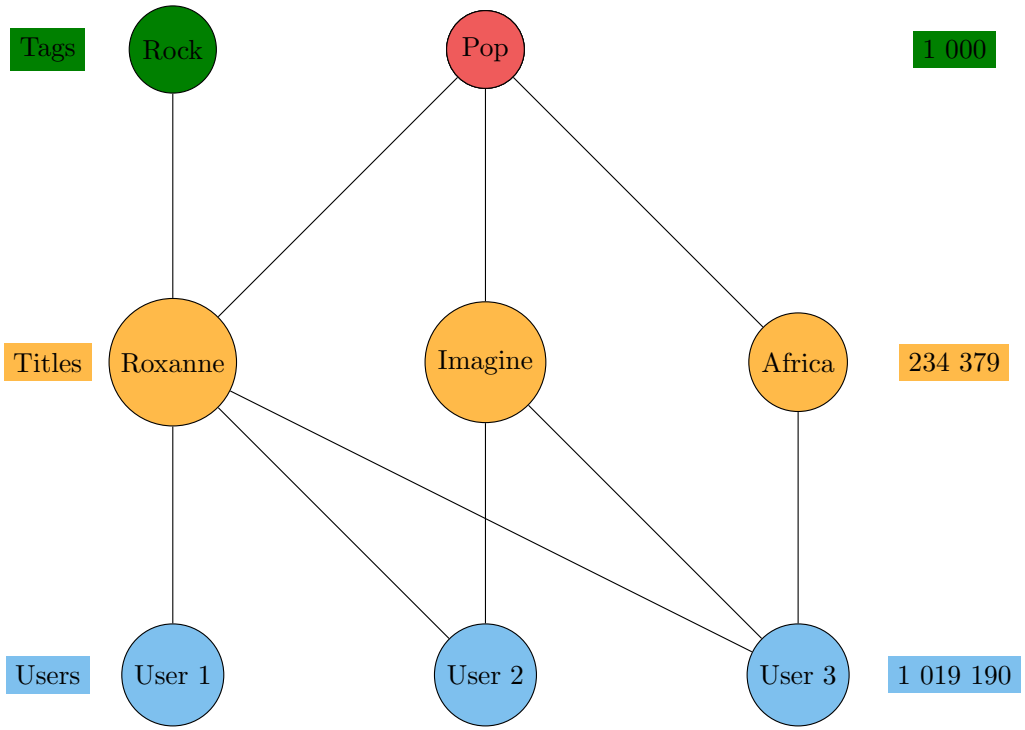
Graphe triparti



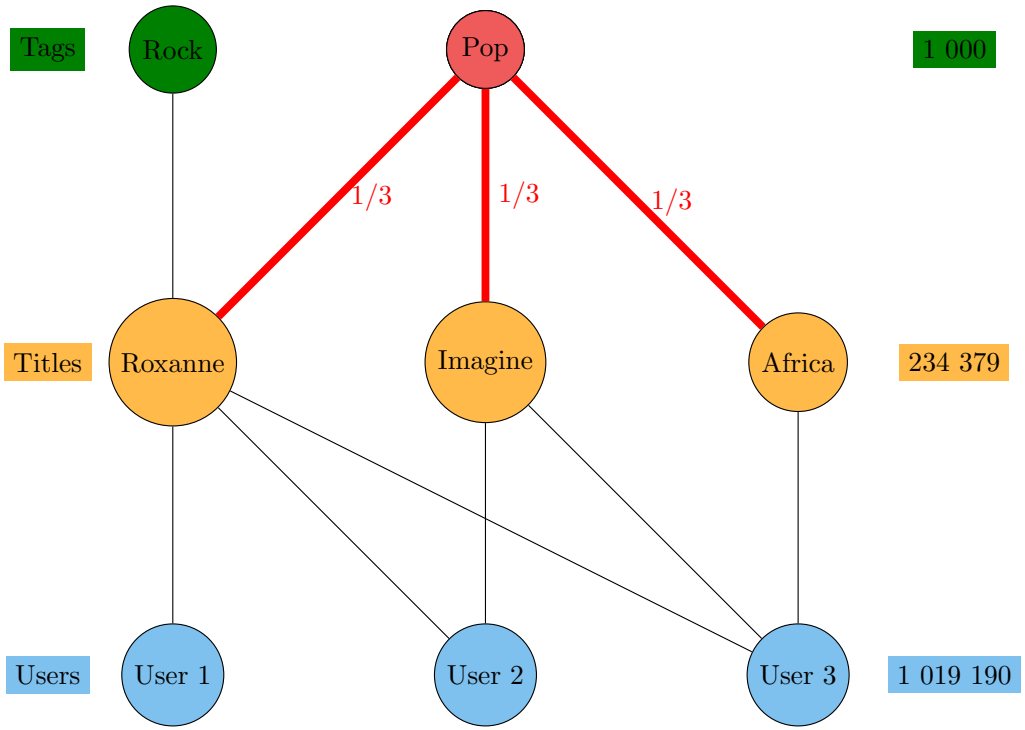
- graphe biparti catégories x produits : $\mathbb{B}_{C\times P}(\mathbb{T}) = (\top, \vdash, E_{\mathbb{T}}^{\top})$
- graphe biparti produits x utilisateurs $\mathbb{B}_{P\times U}(\mathbb{T}) = (\vdash, \perp, E_{\mathbb{T}}^{\perp})$ $(+\ w_{E_{\mathbb{T}}^{\perp}}\ ?)$

Comment mesurer la diversité exprimée dans un graphe triparti ?

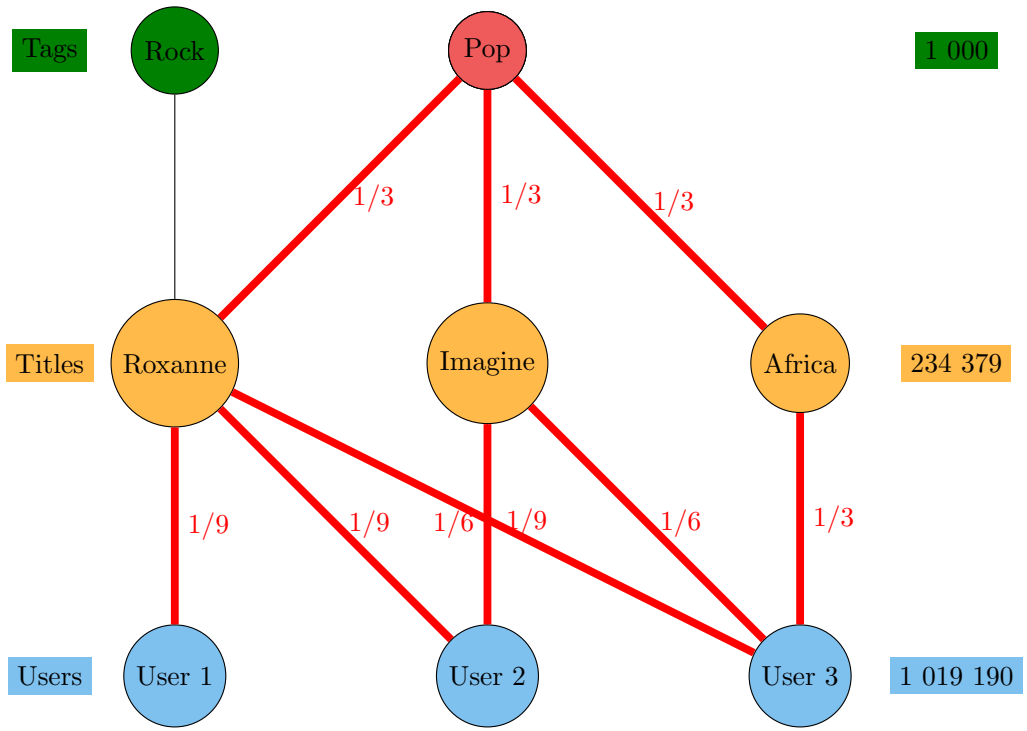
Marche aléatoire



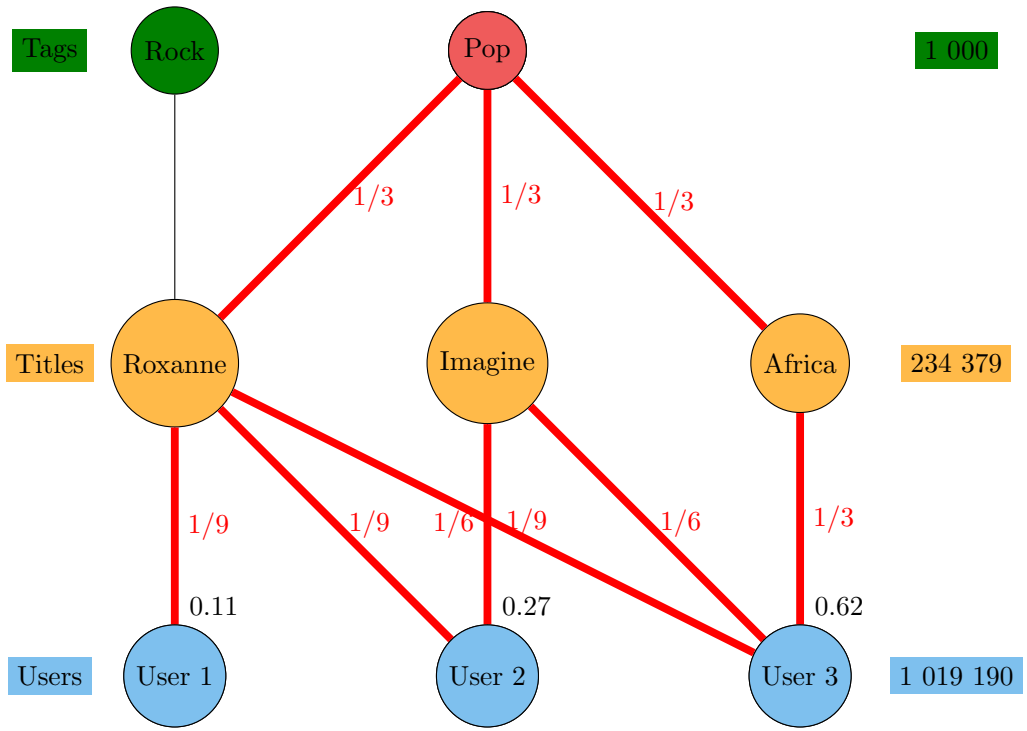
Marche aléatoire



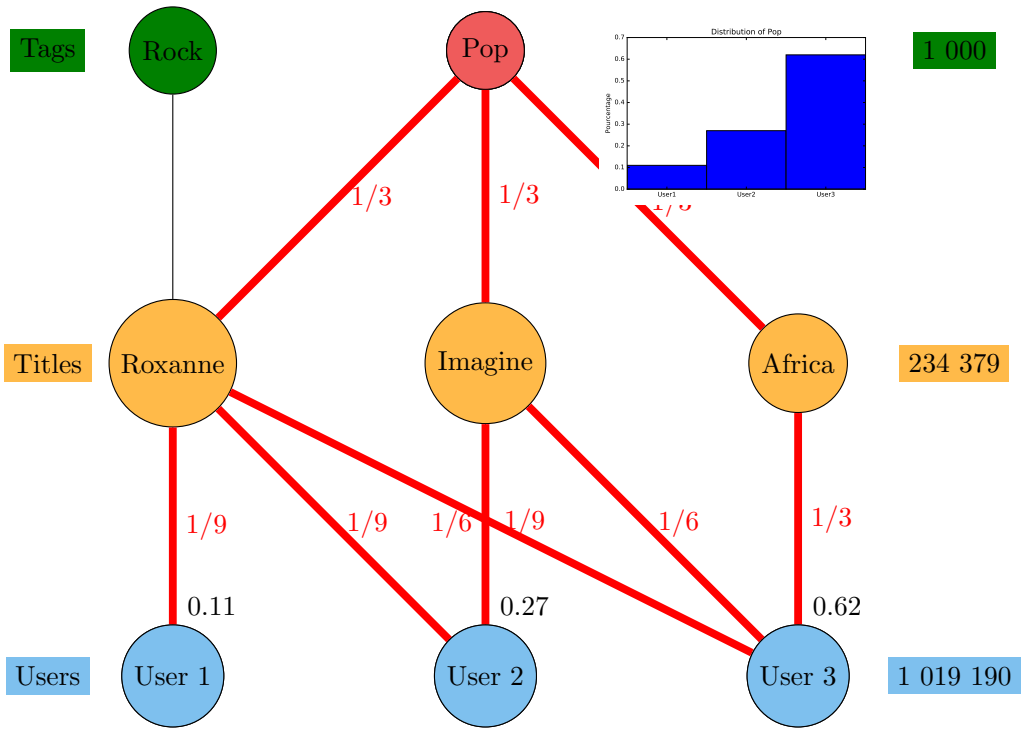
Marche aléatoire



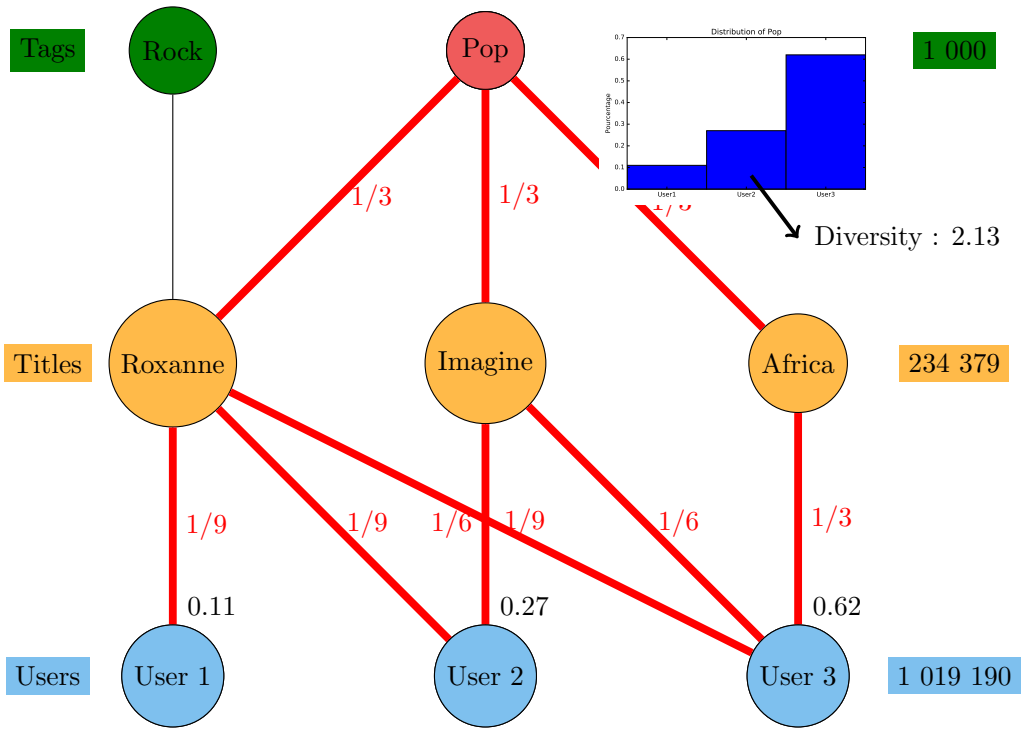
Marche aléatoire



Marche aléatoire



Marche aléatoire

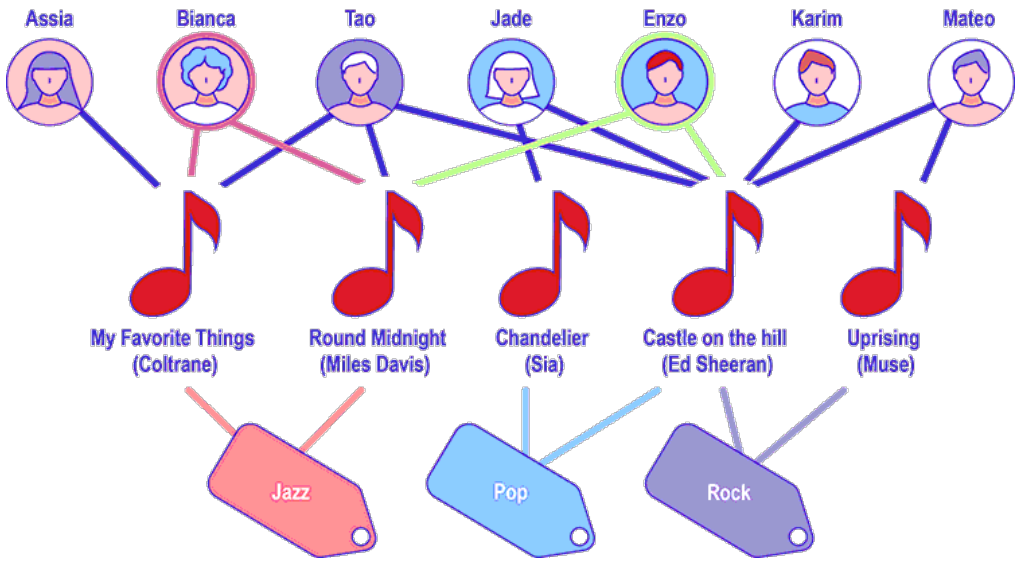


Profils de diversités

Objectif: étudier la diversité du graphe biparti induit par un nœud donné $\sim$ *profil de diversité* d'un nœud.

Proposition: brancher le résultat de marches aléatoires sur des indices classiques de dispersion. Soit $P = \mathrm{RandWalk}(\mathbb{T}, u) = (p_i)_i$:

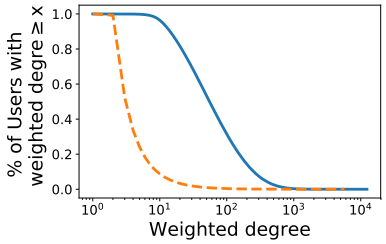
$$\alpha-\text{diveristy} : \mathtt{div}_\alpha(P) = \left(\sum_i p_i^\alpha\right)^{\frac{1}{(1-\alpha)}}$$

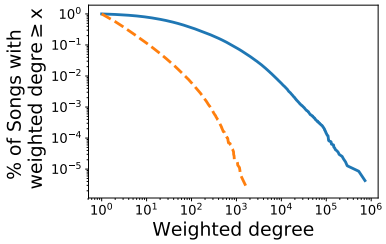


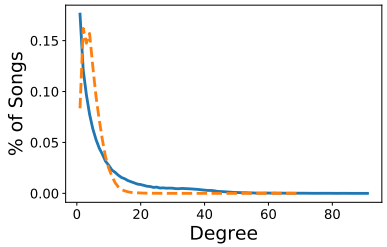
$\mathtt{div}_2(\mathbb{T}, \textit{Bianca}) = 1$ mais $\mathtt{div}_2(\mathbb{T}, \textit{Enzo}) = \frac{8}{3}$

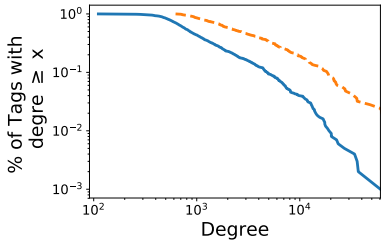
$\mathtt{div}_2(\mathbb{T}, \textit{Jazz}) = 3.6$ mais $\mathtt{div}_2(\mathbb{T}, \textit{Rock}) = 2.5$

Jeux de données









Dataset *MSD*

Issu du projet *Million Song Dataset* (*The Echo Nest* $\mapsto$ *Spotify*) :

- profils d'écoutes : 48 M de triplets ($\sim$ 1M utilisateurs, 300 000 titres)
- catégorisation (*LastFM*) : ($\sim$ 500 000 titres et catégories)

Preprocessing : nettoyage + focus sur les 1000 tags les plus populaires

Dataset *AMZ*

Données *Amazon* (categories *CDs & Vinyl* et *Digital Music*):

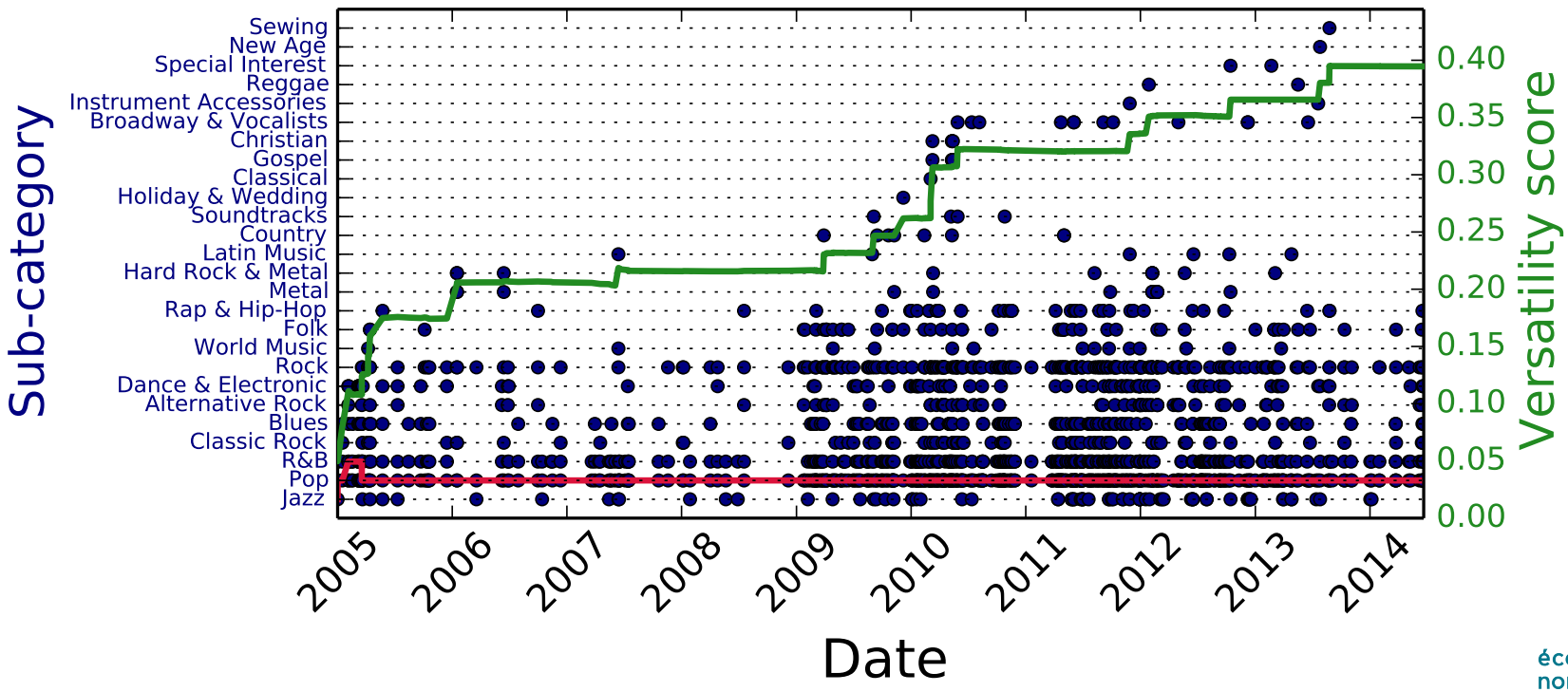
- profils de reviews ($\sim$ 500 000 utilisateurs, 450 000 titres)
- metadata : association titre/catégoires (sous forme arborescente)

Preprocessing : nettoyage + focus sur les 250 catégories les plus utilisées

Un exemple

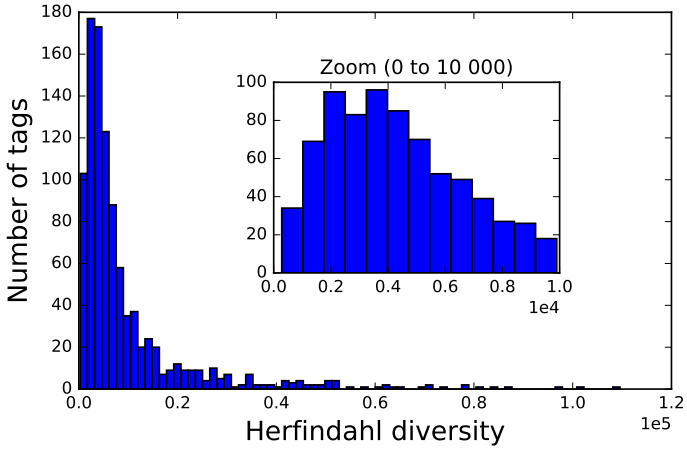
Category Digital_Music (1)
User A3HU0B9XUEVHIM

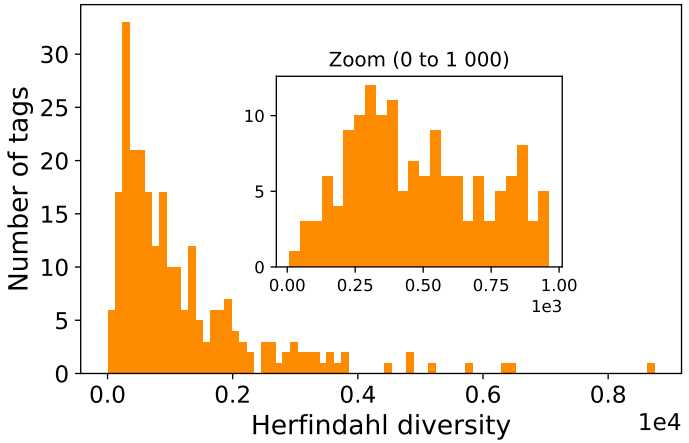
Diversity index: 0.88038
Diffusion index: 0.45
Versatility index: 0.39465



Diversité de l'audience (d'une catégorie)

Diversité (tous les tags)





Distribution hétérogène :

- max $\geq$ 100 000
- moyenne: 9 699 / médiane: 5 111

Homogène sur diversités faibles :

- 75% $\leq$ 10 000
- moyenne/médiane : $\sim$ 4 000

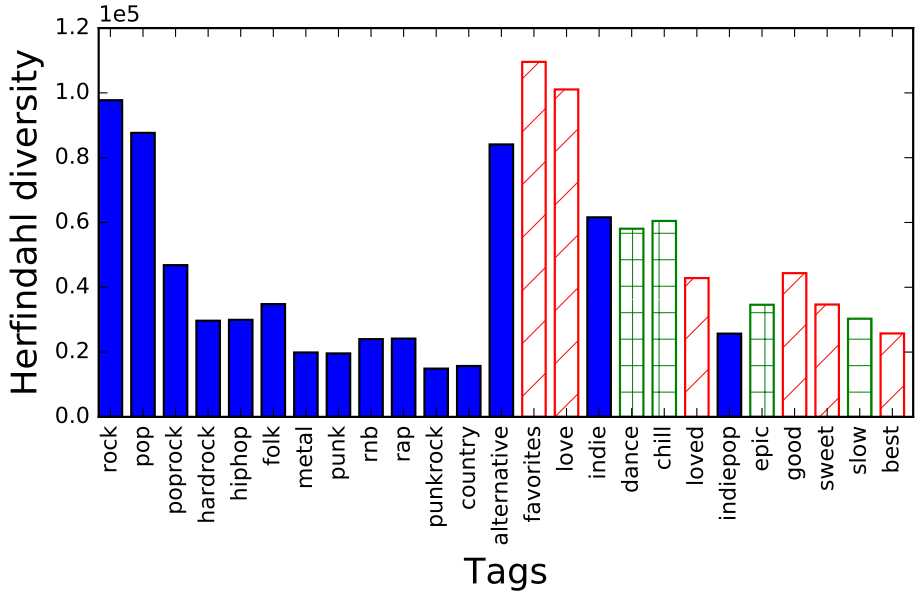
Distribution hétérogène :

- max $\geq$ 10 000
- moyenne: 1 197 / médiane: 806

Homogène sur diversités faibles :

- 60% $\leq$ 1 000
- moyenne/médiane : $\sim$ 500

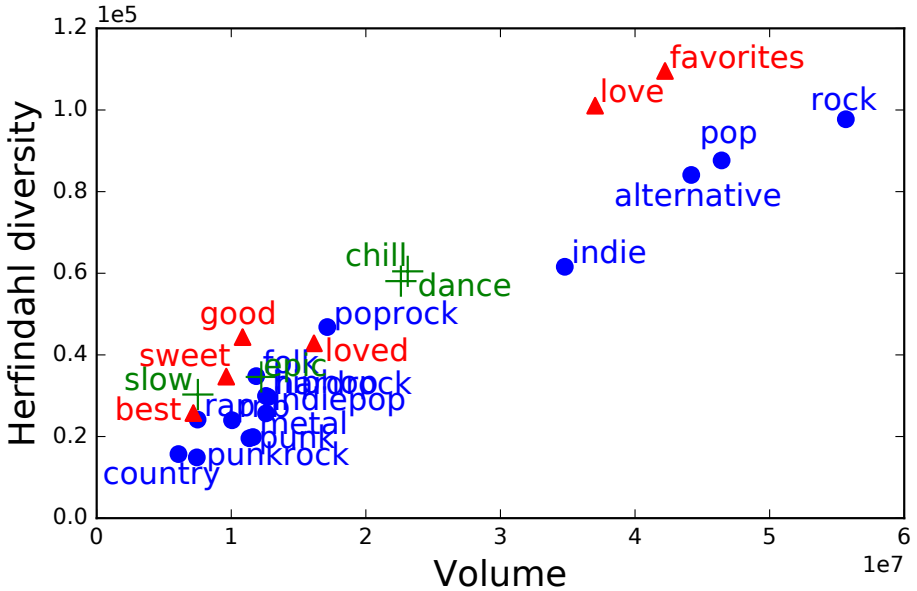
Diversité (25 tags)



Différentes utilisations des tags :

- tags *styles musicaux*
- tags *génériques*
- tags *mixes*

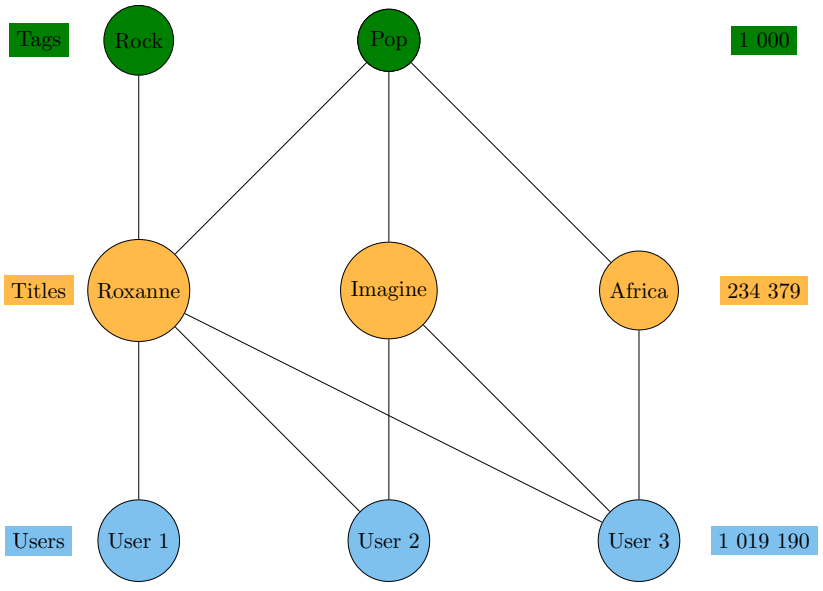
Diversité & volume



- Forte corrélation
- À volume fixe, les génériques et mixes sont plus diversifiés

Comment tenir compte de l'effet du volume de l'audience sur la diversité ?

Normalisation



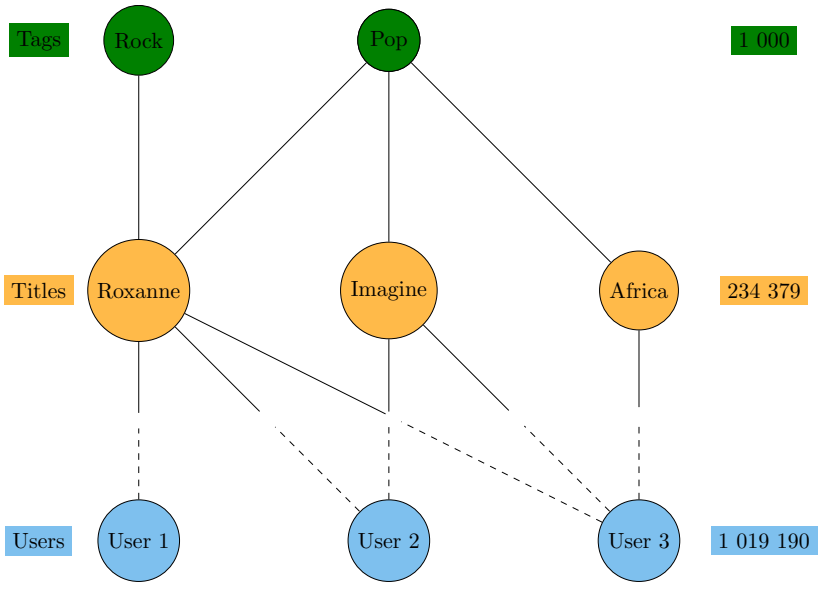
Calibrated herfindahl diversity *(diversité attendue)*

Approche basée sur le *Configuration model* :

- liens du hauts inchangés
- ré-arrangement aléatoirement des les liens du bas.

$$\texttt{chd}_\alpha(\mathbb{T}, u) = \frac{\texttt{div}_\alpha(\mathbb{T}, u)}{\texttt{div}_\alpha(\text{Rand}(\mathbb{T}), u)}$$

Normalisation



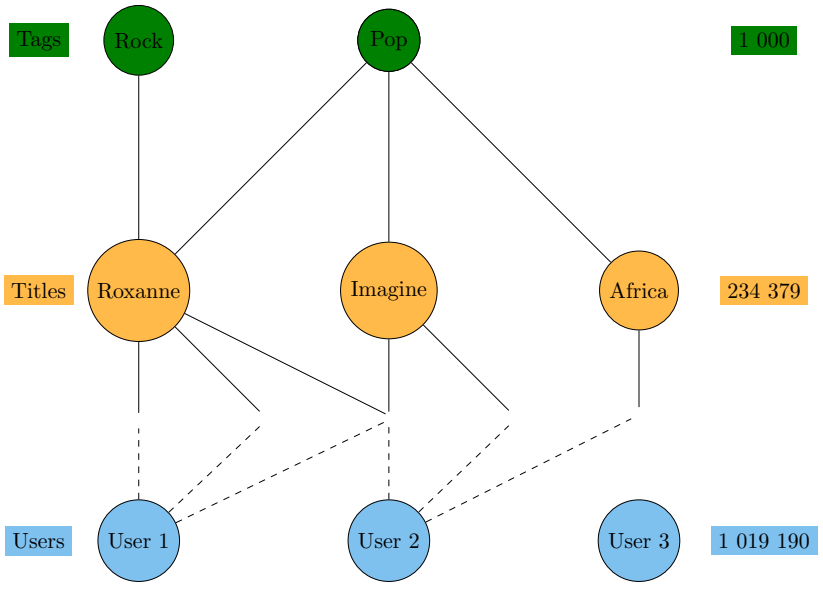
Calibrated herfindahl diversity *(diversité attendue)*

Approche basée sur le *Configuration model* :

- liens du hauts inchangés
- ré-arrangement aléatoirement des les liens du bas.

$$\mathtt{chd}_\alpha(\mathbb{T}, u) = \frac{\mathtt{div}_\alpha(\mathbb{T}, u)}{\mathtt{div}_\alpha(\mathrm{Rand}(\mathbb{T}), u)}$$

Normalisation



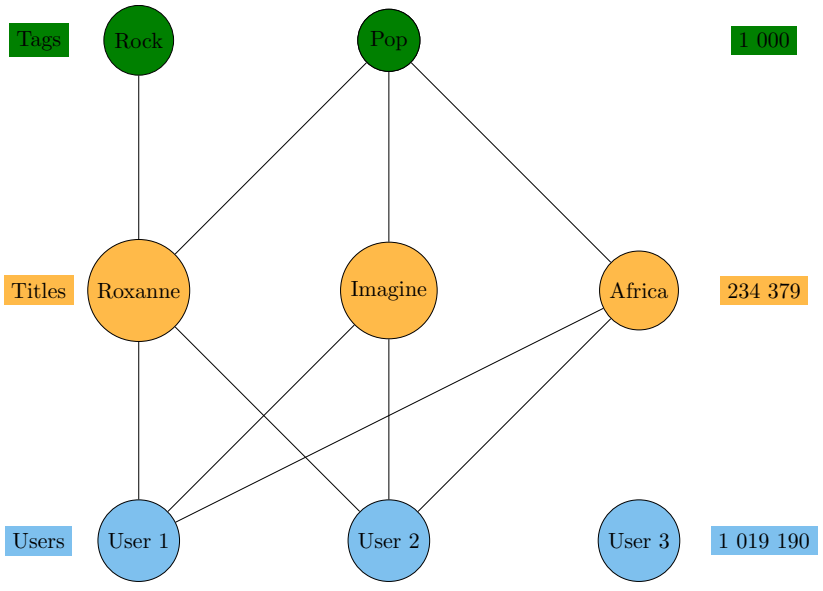
Calibrated herfindahl diversity *(diversité attendue)*

Approche basée sur le *Configuration model* :

- liens du hauts inchangés
- ré-arrangement aléatoirement des les liens du bas.

$$\texttt{chd}_\alpha(\mathbb{T}, u) = \frac{\texttt{div}_\alpha(\mathbb{T}, u)}{\texttt{div}_\alpha(\text{Rand}(\mathbb{T}), u)}$$

Normalisation



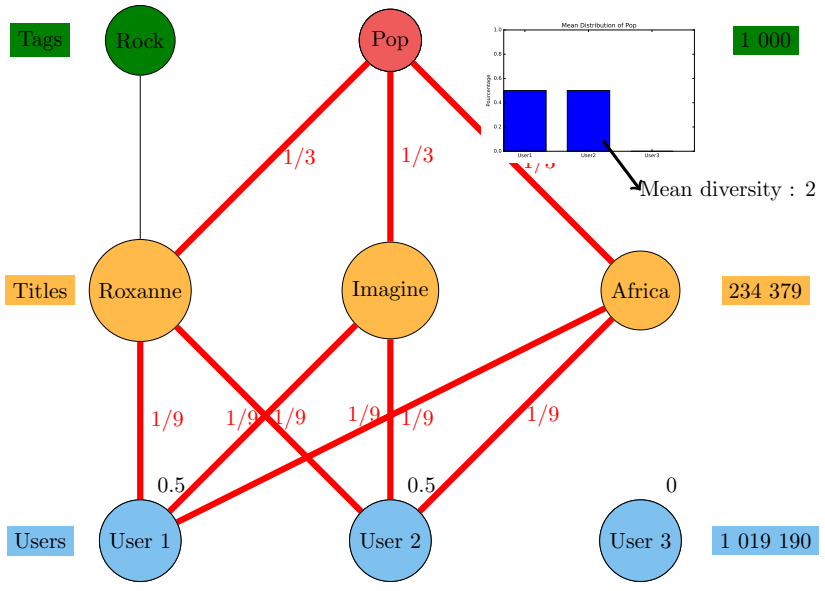
Calibrated herfindahl diversity *(diversité attendue)*

Approche basée sur le *Configuration model* :

- liens du hauts inchangés
- ré-arrangement aléatoirement des les liens du bas.

$$\texttt{chd}_\alpha(\mathbb{T}, u) = \frac{\texttt{div}_\alpha(\mathbb{T}, u)}{\texttt{div}_\alpha(\text{Rand}(\mathbb{T}), u)}$$

Normalisation



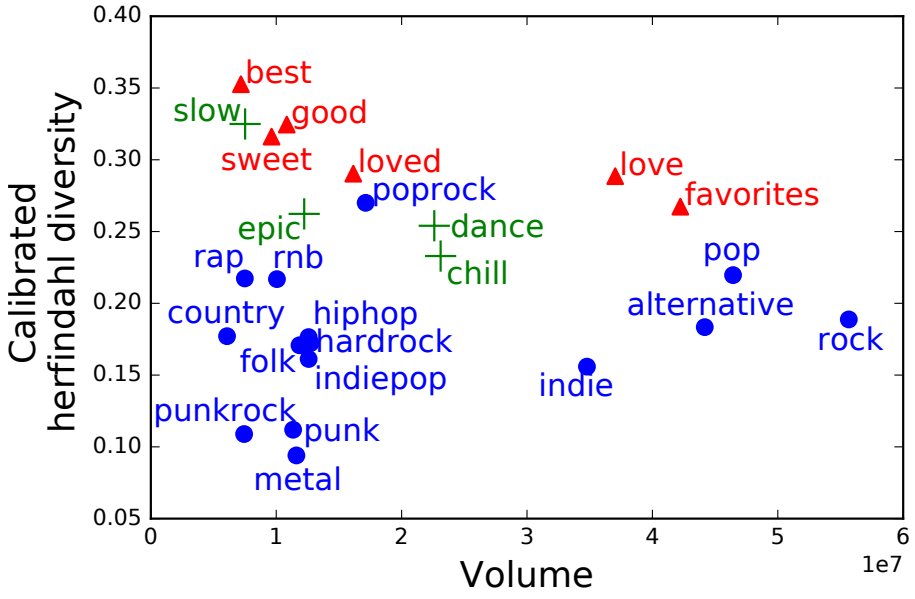
Calibrated herfindahl diversity *(diversité attendue)*

Approche basée sur le *Configuration model* :

- liens du hauts inchangés
- ré-arrangement aléatoirement des les liens du bas.

$$\mathtt{chd}_\alpha(\mathbb{T}, u) = \frac{\mathtt{div}_\alpha(\mathbb{T}, u)}{\mathtt{div}_\alpha(\mathrm{Rand}(\mathbb{T}), u)}$$

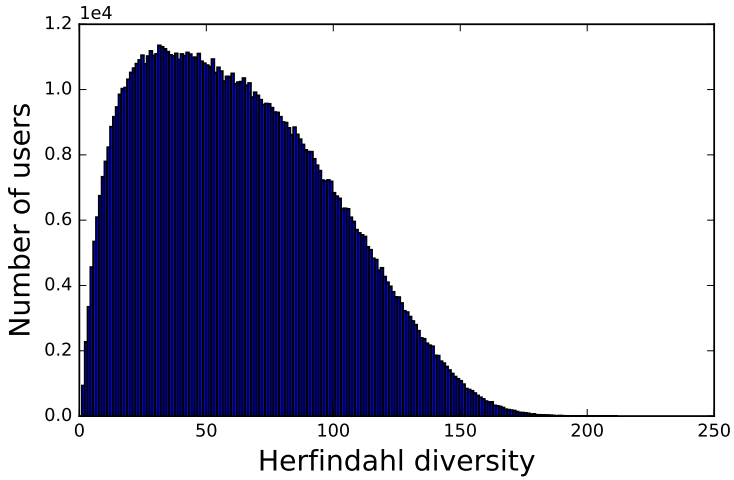
Diversité normalisée & volume

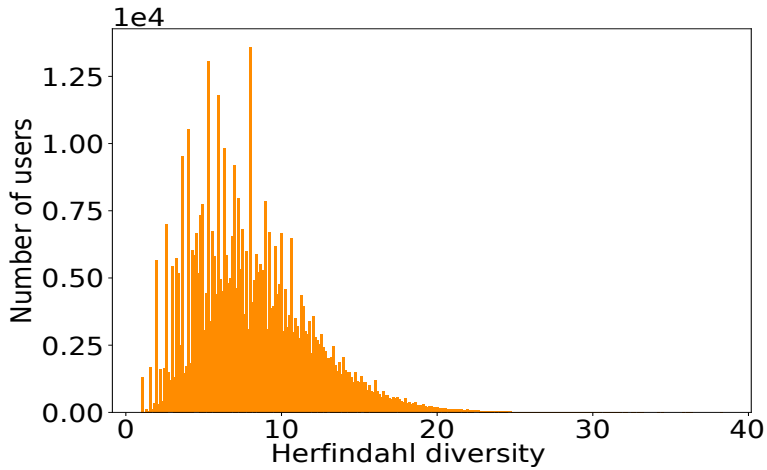


- Plus de corrélation forte (mais accroissement tout de même)
- *love* et *best* deviennent comparables à *favorite*
- *metal*, *punk*, ... sont identifiés comme des niches musicales

Diversité de l'attention (d'un utilisateur)

Diversité de l'attention

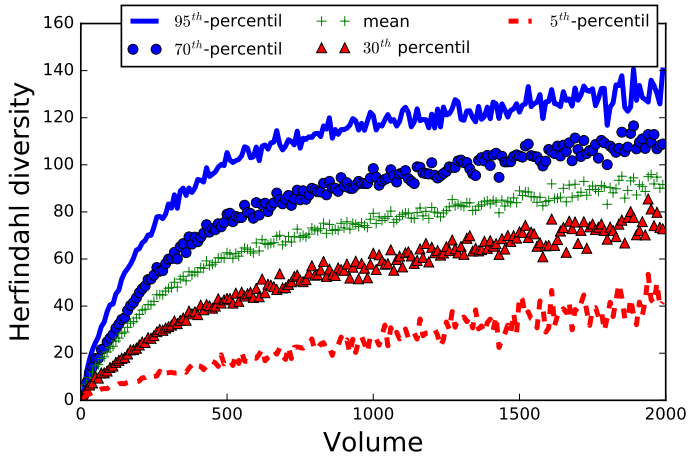


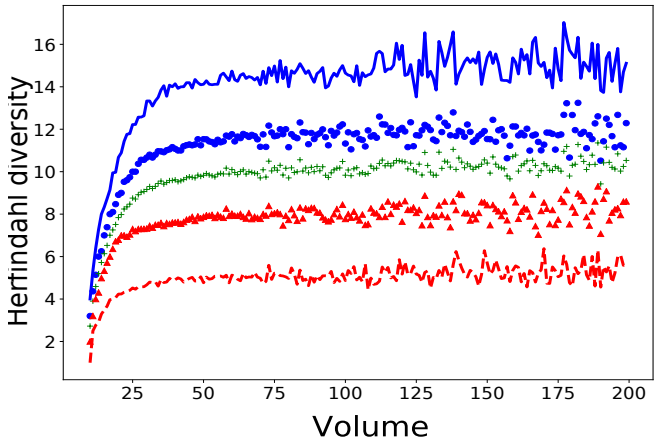


- Distribution plutôt homogène
- médiane à 59 (max th = 1 000)

- Distribution plutôt homogène
- Médiane à 7 (max th = 250)

Diversité vs. volume





- 85% des utilisateurs ont un volume d'écoute entre 10 et 2 000
- seuil de saturation : $\sim$ 500

- 96% des utilisateurs écrivent moins de 100 commentaires
- seuil de saturation : $\sim$ 40

Mise en évidence d'un phénomène de saturation

Modèles aléatoires

Reformulation

Nouvelle formulation du problème

Si on oublie un instant l'aspect graphe, on peut se ramener au problème suivant : soit

- ν : nb d'objects (*utilisateurs*)
- τ : nb d'ensembles (*catégories*)
- v_u : la valeur de l'objet $u < \nu$ (*nb de fois qu'un titre est écouté par u*))
- a_t : la probabilité de placer un objet dans l'ensemble t

L'espérence E liée à Herfindahl est donnée par :

$$E = \sum_t a_t (\sum_u {v_u}^2 + a_t (1 - \sum_u v_u^2))$$

$\Longrightarrow$ Différents modèles donnent différentes espérences.

Différents modèles

baseline: un utilisateur u choisit un tag avec égale probabilité ($a_t = \frac{1}{\tau}$ et $v_u = \frac{1}{V_u}$, où V_u est le volume de u)

pas de couche du milieu (titres)

song-uniform: un utilisateur u choisit aléatoirement (uniformément) un titre s. u répartit sa valeur v_u en $\frac{1}{d_\top(s)}$ part égale placées dans les $d_\top(s)$ tags reliés à s

c'est le *calibrated herfindahl diversity*

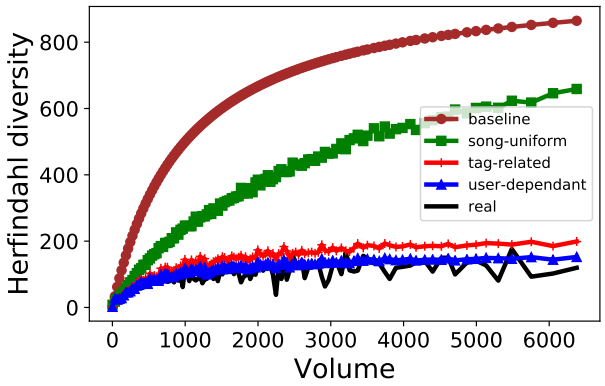
tag-related: On ajoute la contrainte que le a_t est proportionel à sa popularité du style de musique t ($a_t = \frac{d_\perp(t)}{\sum_{t'} d_\perp(t')}$).

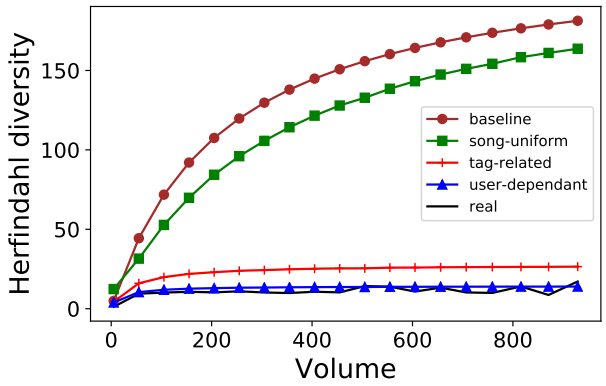
impact de la popularité des styles

user-dependant: On rajoute le fait que les choix des titres ne sont plus des variables indépendantes.

impact du comportement répétitif des utilisateurs

Des contraintes réalistes





- Même phénomène de saturation quelque soit le modèle
- Plus les contraintes sont réalistes, plus on s'approche du comportement empirique
- *tag-related* et *user-dependant* sont particulièrement proches des mesures empiriques
- impact de *tag-related* : facteurs exogènes (popularité hétérogène des styles de musique) qui expliquent les limites de la diversité empirique

Systèmes de recommandation

Une expérience

Filtrage collaboratif pour données implicites

- Utilisé par bcp de plateformes (*Netflix*, *Youtube*, *Spotify*, *Amazon*, ...)
- Fonder une recommandation sur ce que des utilisateurs *similaires* ont aimé
- S'appuie sur la *Factorisation Matricielle* (Matrice = Utilisateurs $\times$ Produits)

Expérience

Après entraintement du modèle sur une partie du jeu de données :

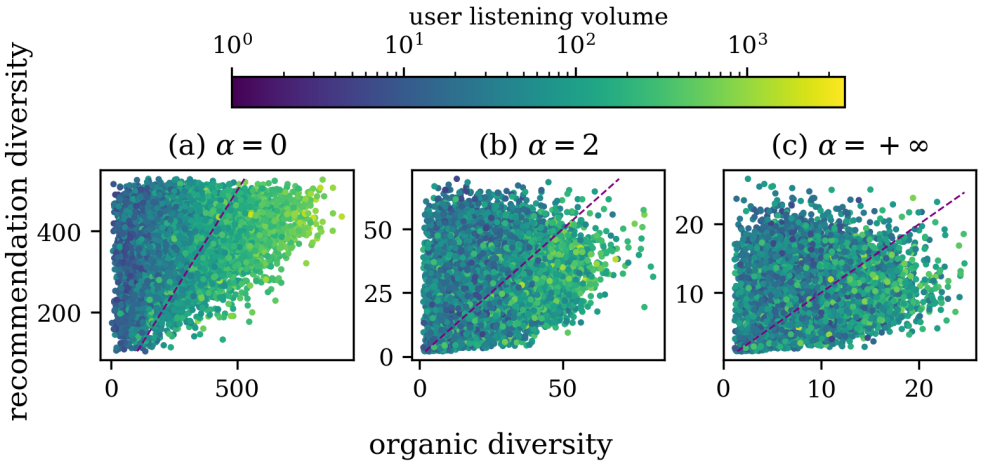
1. Prédiction: inférence de la *force* des liens manquants entre utilisateurs et produits
2. Sélection: pour chaque utilisateur, classement puis sélection des k-top produits ($k = 10, 50, 500$)

Ce qui permet de mesurer :

- la diversité des utilisateurs avant recommandation
- la diversité des utilisateurs après avoir été exposé à k recommandations (décroissance linéaire des poids)

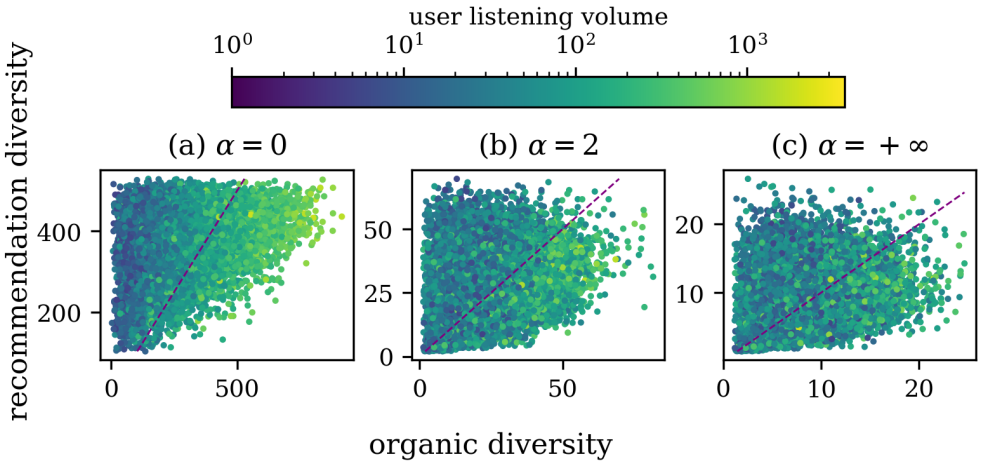
Diversité des recommdantions

Est-elle liée à celle des utilisateurs?



Diversité des recommdantions

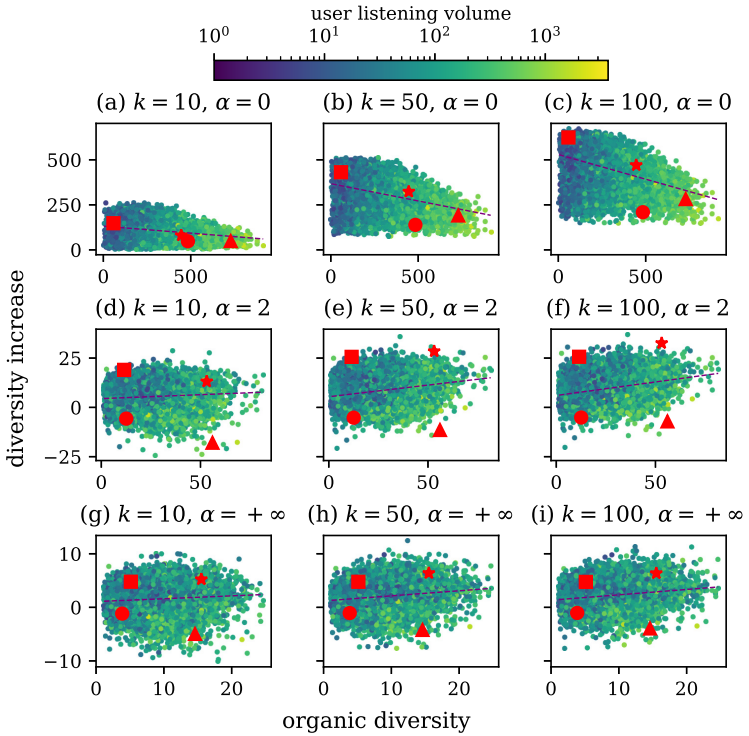
Est-elle liée à celle des utilisateurs?



- pas de correlation forte entre la diversité organique et celle des recommandations
- les recommandations tendent à être plus diversifiées que le profil d'écoute

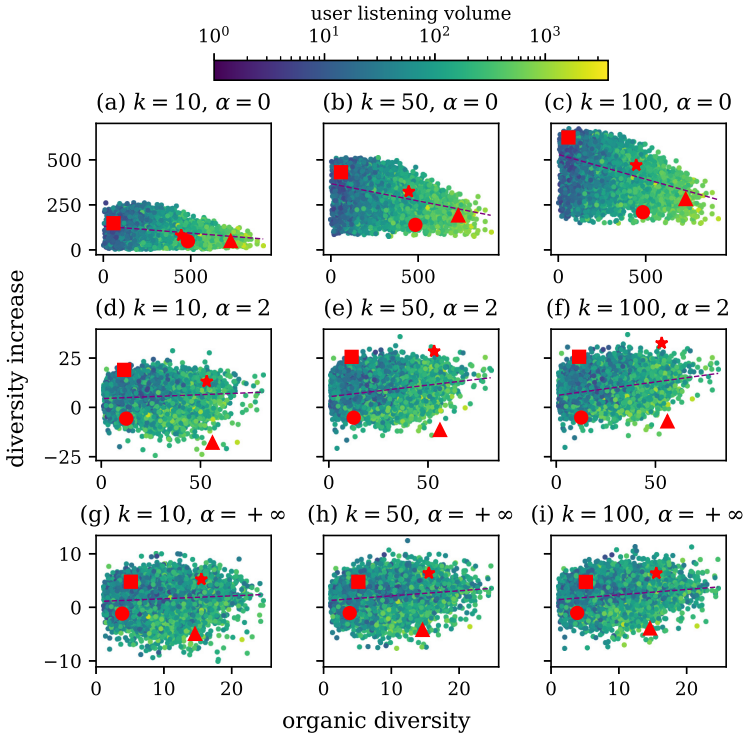
Effets des recommandations

Quelle diversité des utilisateurs
après avoir été exposés aux recommandations ?



Effets des recommandations

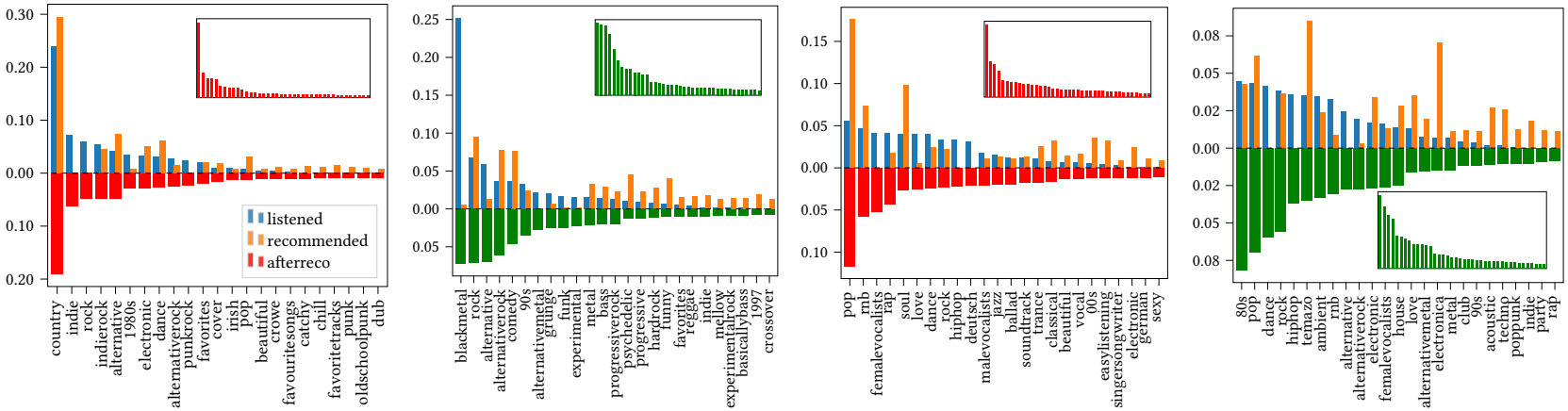
Quelle diversité des utilisateurs
après avoir été exposés aux recommandations ?



- La majorité des utilisateurs voit sa diversité augmentée par les recommandations
- La diversité augmente avec le nombre de produits recommandés
- L'effet sur la diversité dépend du profil de l'utilisateur

Différents effets

Est-il possible de savoir quel va être l'effet des recommandation
étant donné un profil d'écoute ?



- Les recommandations augmentent la variété des produits
- mais échouent à fournir une exposition équilibrée

Conclusions

Pour dériver une valeur de α-diversité, nous avons besoin de :

- un graphe multi-parti (au moins triparti)
- un ordre de diversité (paramètre α)
- un méta-chemin afin de dériver une marche aléatoire (constrainte)

Ce qu'on a montré (sur 2 datasets)

1. saturation de la diversité de l'attention des utilisateurs — **Analyse**
2. le manque de diversité provient de facteurs exogènes — **Modélisation**

Investigating the lack of diversity in user behavior: The case of musical content on online platforms. Rémy Poulain and Fabien Tarissan, *Information Processing & Management*, 57(2), Eslevier, 2020.

Fondements théorisques (axiomatisation)

Measuring Diversity in Heterogeneous Information Networks. Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, Fabien Tarissan. In *Theoretical Computer Science*, Elvesier, 2021

Analyse de l'impact des recommandations sur la diversité

- les recommandations améliorent la diversité de la plupart des utilisateurs
- mais échouent à générer une exposition équilibrée des styles musicaux

Measuring the effect of collaborative filtering on the diversity of users' attention. Augustin Godinot, Fabien Tarissan. In *Applied Network Science*, 8(1):9, Springer, 2023.

Questions?

http://tarissan.complexnetworks.fr/