Une analyse de l'éco-système YouTube: le réseau des commentateurs.

Annick Vignes

CAMS-EHESS, Ecole des Ponts and INRAE

Ecole des Chartes, le 18 janvier 2024.

YouTube, un socio-éco-système?

Depuis 2005, Youtube est la plus importante des plateformes numériques. Y interagissent, des vidéastes et des "spectateurs/ commentateurs". Sa valeur économique se mesure à l'aune de ses interactions.

Questions de recherche: quel est son modèle économique? Peut-on parler d'un marché Youtube? Quels seraient les acteurs de ce marché?

Multiples modes de financement : publicité, abonnements, crowdfunding....
Youtube entre marché biface et logique de don/ contre-don?

ANR: (APY) Entre Réseaux Complexes et Marché : YouTube au prisme des Sciences sociales computationnelles

Travail en collaboration avec Kurt Kusterer et Sylvain Mignot.

Partenariat avec une entreprise, Wizdeo, qui collecte des données Youtube

A partir des données fournies par cette entreprise, le projet se propose d'analyser les modes de financement de la plateforme, entre logique de marché (plateforme, branding, abonnements) et logique de dons/ contre-dons (crowdfunding)

Projet interdisciplinaire entre des computer scientists, des économistes et des sociologues.

Les données et leurs limites.

Une base de données sur une année (2020):

- **Une base de données de 36 973 channels (vidéastes)**
  - couvre le paysage du YouTube français
  - pour chacun de ces channels, on connait l'identité du vidéaste, son champ thématique (14 champs thématiques répertoriés) et ses revenus

- **Une base de données de 391 588 vidéos.**
  - Pour chaque vidéo, on connait le nombre de like, de dislike, le temps de visionnage, le channel qui l'a produit.

- **Une base de données de 12 012 966 commentaires.**
  - Pour chaque commentaire, on connait le nom du commentateur, la vidéo commentée et le moment de téléchargement.
  - On ne connait pas la teneur des commentaires: pas d'analyse textuelle possible.

Les données manquantes

La production de premières statistiques descriptives révèlent que :

- Seuls 71 channels (0.2% de l'échantillon) contiennent plus de 80% des informations. Les revenus s'étalent entre 0 et 2 millions d'€.
- Les revenus sont connus pour 59 969 videos, soit 15 % du data set.
- Le revenu moyen est de 134 € par video, standard deviation de 605 et le revenu max. de 73 134 €. **La limite inférieure du quartile supérieur est de 93 €.**

Distribution des revenus dans notre échantillon

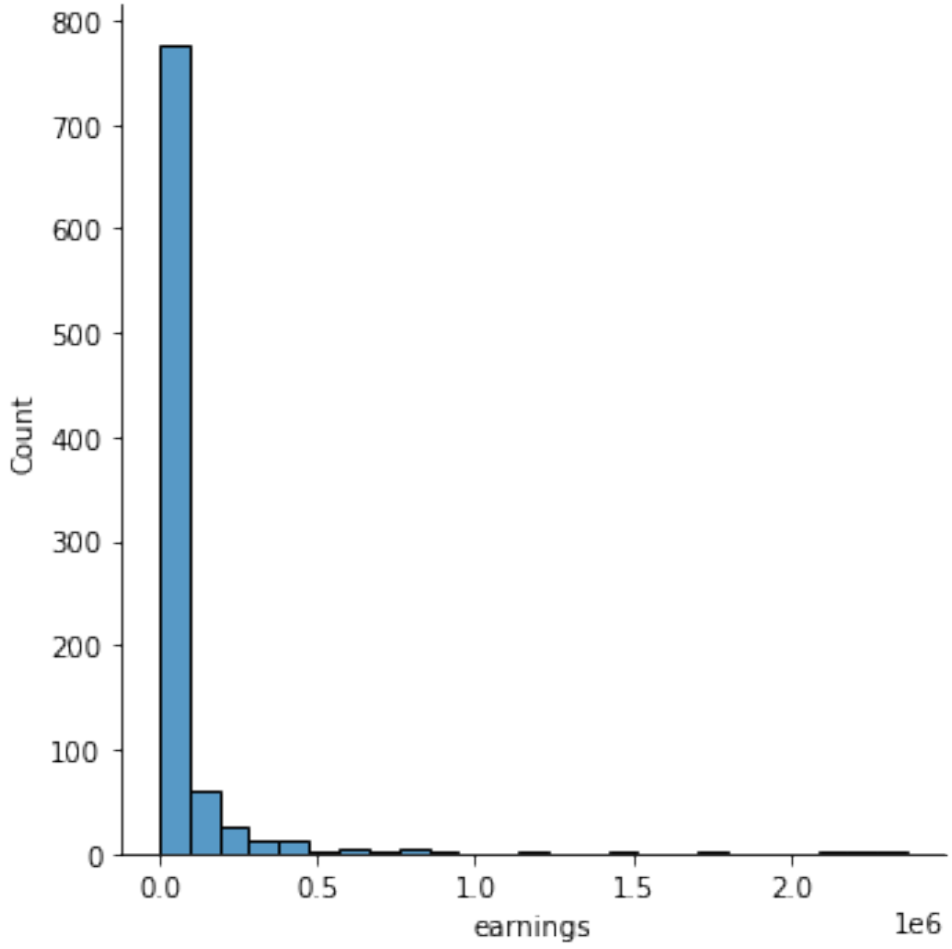


Figure: Distribution of earnings (X-axis to be read as: $x.10^6$)

Que faire?

1. il nous manque les données dont on avait besoin
2. les revenus qu'on connait suivent une loi puissance. La moyenne ne signifie rien.
3. aucune indication sur la représentativité de cet échantillon.

Le fonctionnement de Youtube

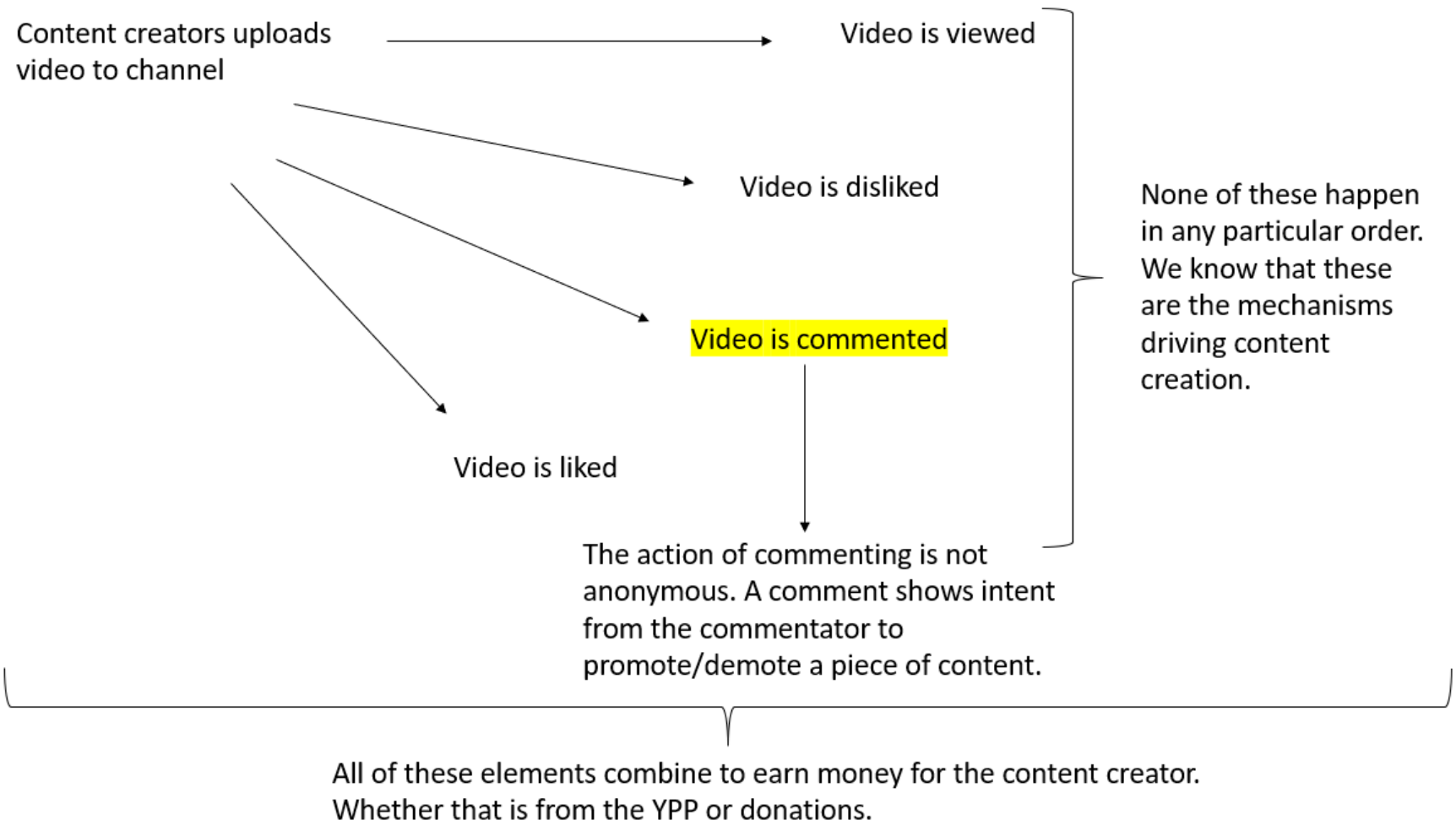


Figure: ?The elements which make up YouTube?

Comment mesurer la visibilité des Youtubeurs?

On sait que:

1. Les vues, les abonnés, les commentateurs contribuent à la popularité des Youtubeurs (channels).
2. Les Youtubeurs les plus populaires sont rétribués par la publicité.
3. Certains d'entre eux en font une profession
4. Le YouTube Partner Program (YPP) organise la rétribution des Youtubeurs selon des règles précises: il faut avoir au moins 1 000 abonnés et 4 000 heures de visionnage de vidéos publiques au cours des 12 derniers mois.

Nous faisons l'hypothèse que:

- les commentateurs influencent la popularité d'une vidéo (et donc sa possibilité de remporter des gains.
- les commentateurs commentent stratégiquement.

LE BLOC - ABILO (Épisode Final..)

Hassan ✓
3,14 M d'abonnés

S'abonner

56 k | Partager

1,4 M de vues il y a 7 mois
Abonne toi la famille, le bloc tous les dimanches !

Écriture de la série : Hassan **Plus**

5 097 commentaires **Trier par**

A Ajoutez un commentaire...

Épinglé par Hassan

Hassan ✓ il y a 7 mois

Qui est là depuis là début de la série ? 🔥

10 k

500 réponses

Réduction de l'échantillon

Objectif: conserver les channels qui répondent aux conditions du YPP: plus de 1000 abonnés et plus de 4000 heures de visionnage.

Les channels avec plus de 1000 abonnés: on passe de 36 973 à 27 343 channels.
Conserver les channels avec plus de 4000 heures de visionnage s'avère plus compliqué car *le temps de visionnage global est très mal informé (moins de 10% des données)*.

Mesurer le temps

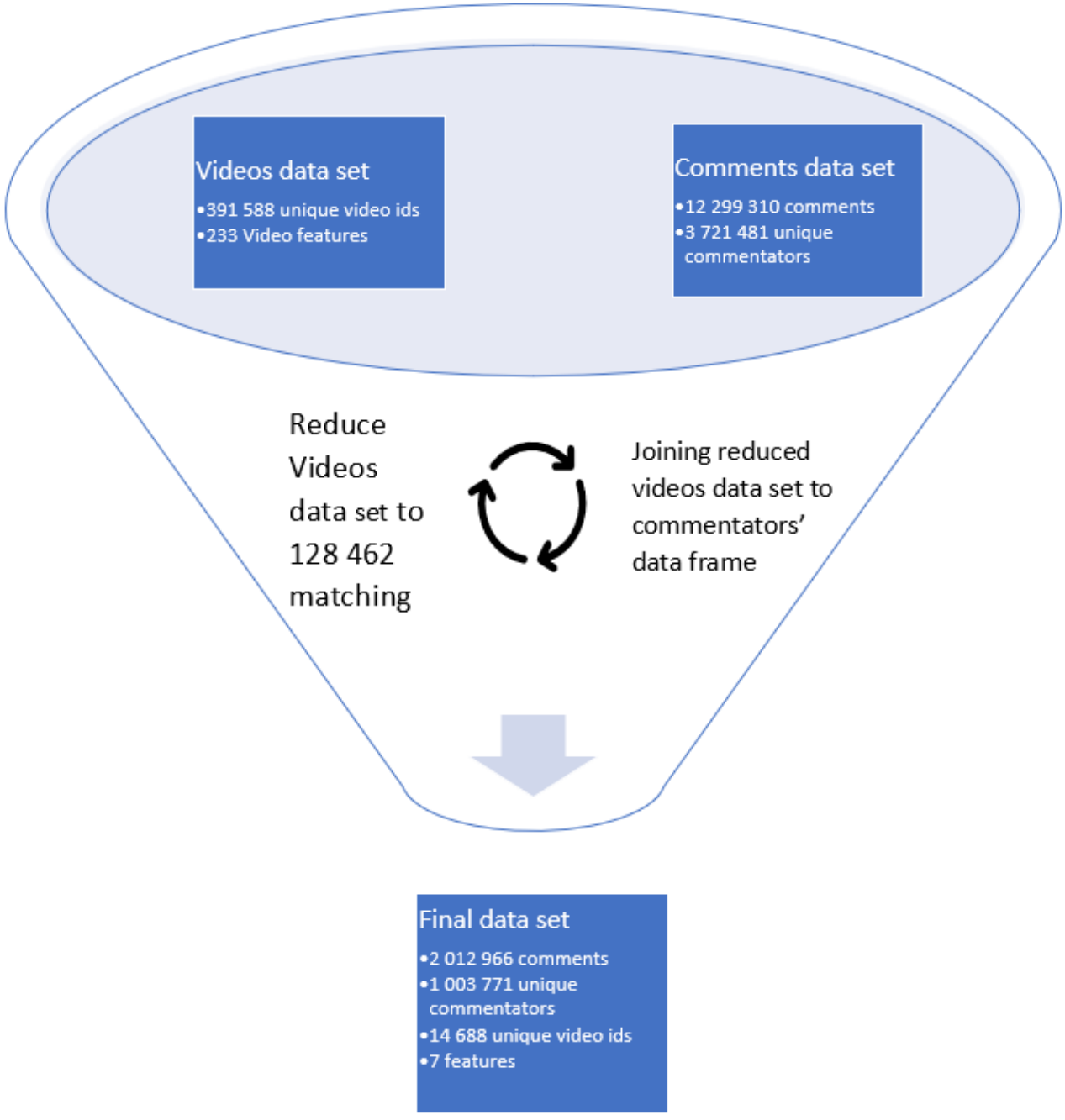
- il faut utiliser le data set des vidéos (qui contient les temps de visionnage)
- il faut calculer le temps de visionnage sur un an, mais on n'a pas plus que six mois
  - Après avoir vérifié que entre le 2ème mois et le 6ème mois, le temps visionné n'augmentait que de 11%, nous prenons le temps à six mois comme proxy du temps à un an.

$\longrightarrow$ Sur 391 495 videos, on garde 128 462 videos.
Intersection des deux sous-ensembles (channels et vidéos):
$\longrightarrow$ Vidéos produites par 4093 channels (parmi les 27343)

La base des commentaires

Avec des techniques identiques aux précédentes, nous allons fusionner la base des vidéos retenues à la base des commentaires. Nous obtenons à la fin une base de commentateurs, les vidéos et channels (et leurs caractéristiques.

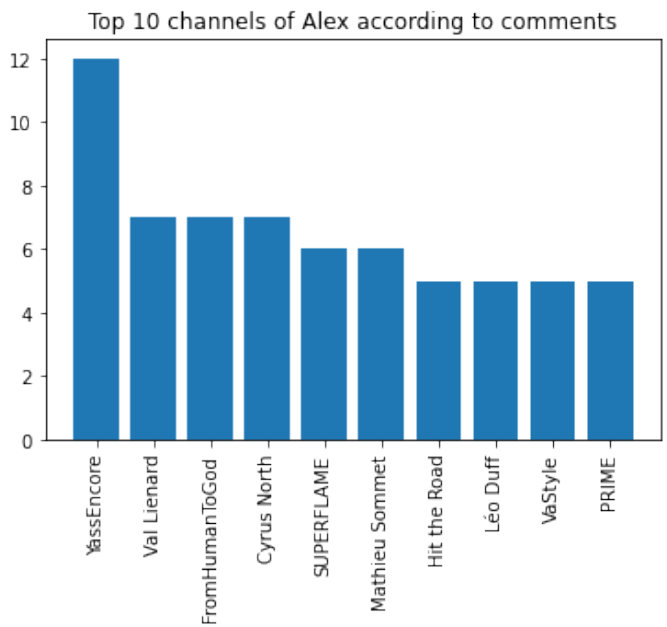
$\longrightarrow$ Comment caractériser leur comportement (stratégique ou aléatoire)?

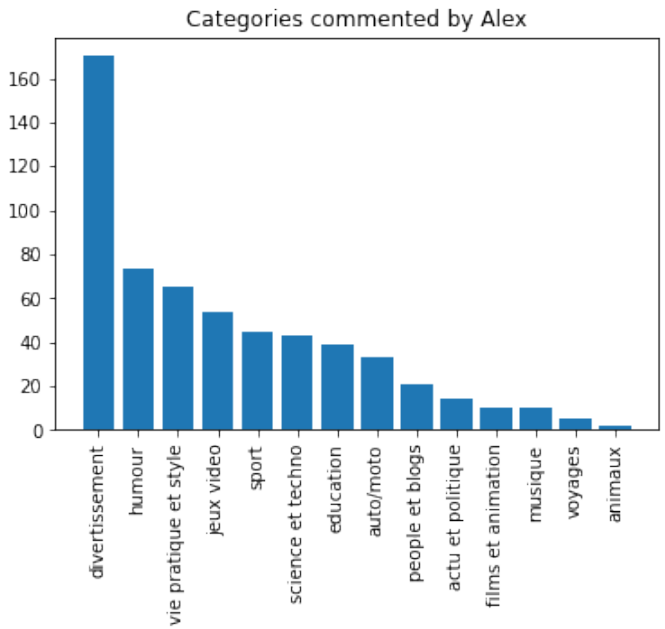

Videos data set
•391 588 unique video ids
•233 Video features
Comments data set
•12 299 310 comments
•3 721 481 unique commentators
Reduce Videos data set to 128 462 matching
Joining reduced videos data set to commentators' data frame
Final data set
•2 012 966 comments
•1 003 771 unique commentators
•14 688 unique video ids
•7 features

Alex, influenceur.

Un des 10 commentateurs les plus importants de notre base de données.

Question: Les stratégies de commentaires sont-elles ciblées sur une ou deux thématique (stratégie de spécialisation) ou diversifiées (assurer une visibilité max.)?





Les 100 commentateurs les plus influents

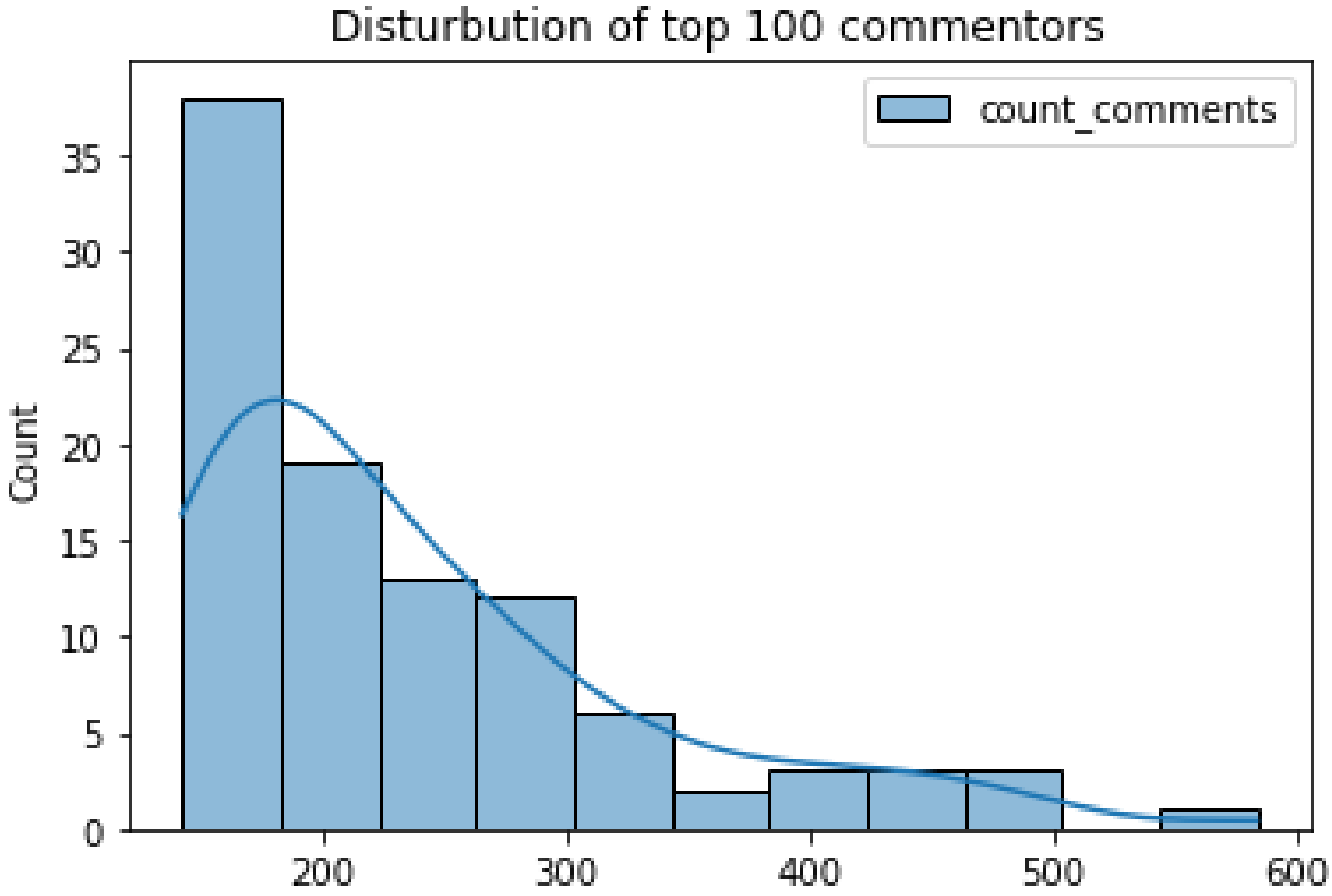
Quelques chiffres:

Nombre moyen de commentaires faits par le top 10% : 15.21

Nbre moyen de commentaires faits par les 90% restants : 1.72

Nombre médian de commentaires faits par le top 10%: 11

Nbre médian de commentaires faits par les 90% suivants: 1


Disturbution of top 100 commentors
count_comments
Count
0
5
10
15
20
25
30
35
200
300
400
500
600

Les commentateurs sont-ils aussi des YouTubers?

Pour les 100 principaux commentateurs, nous vérifions leur profil à la main.
Un YouTuber français est une chaîne dont l'origine est la France et qui compte au moins 1000 abonnés.

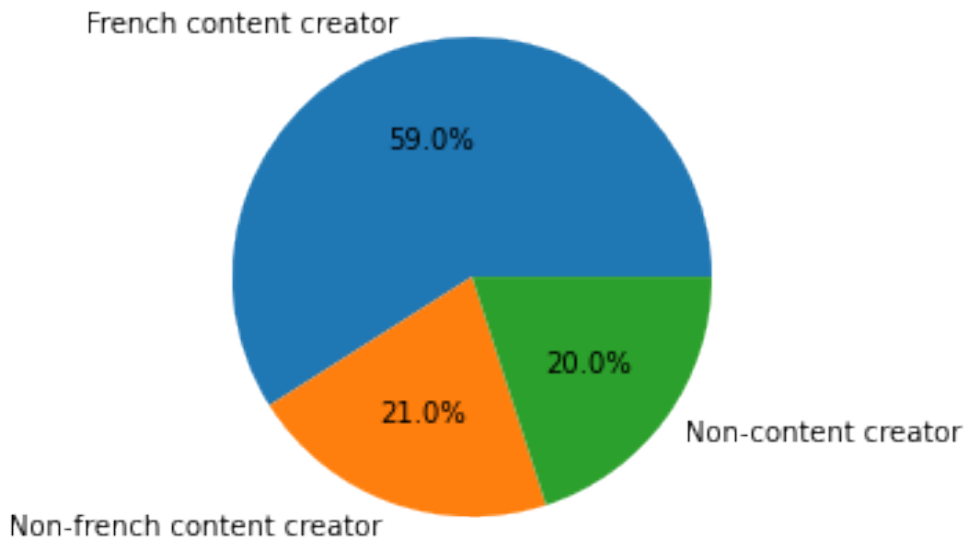
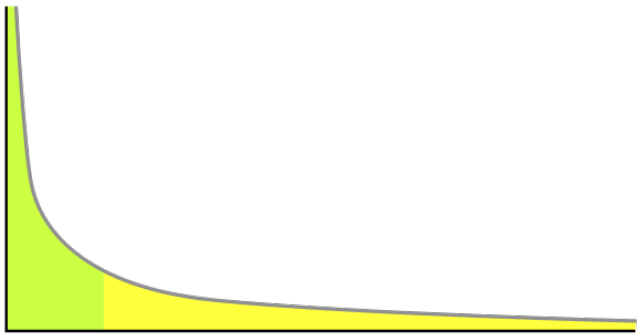


Figure: Top 100 commentators by origin

Loi puissance



source:wikipedia

La distribution d'une loi de puissance, correspondant à un classement de popularité des sites web. À gauche la zone verte illustre le principe des 80-20 du principe de Pareto. À droite la queue de la distribution illustre l'effet longue traîne.

Sur Internet, moins de 1 % de la population contribue de façon active, 9 % participe occasionnellement et 90 % sont des consommateurs passifs, qui ne contribuent jamais.

Une loi de puissance est une loi qui peut s'écrire: $y = ax^k$

Sur un graphique aux échelles logarithmiques, le graphe d'une loi de puissance est une droite. En effet, la relation ci-dessus peut s'écrire :

$\log(y)=k\log(x)+\log(a)$

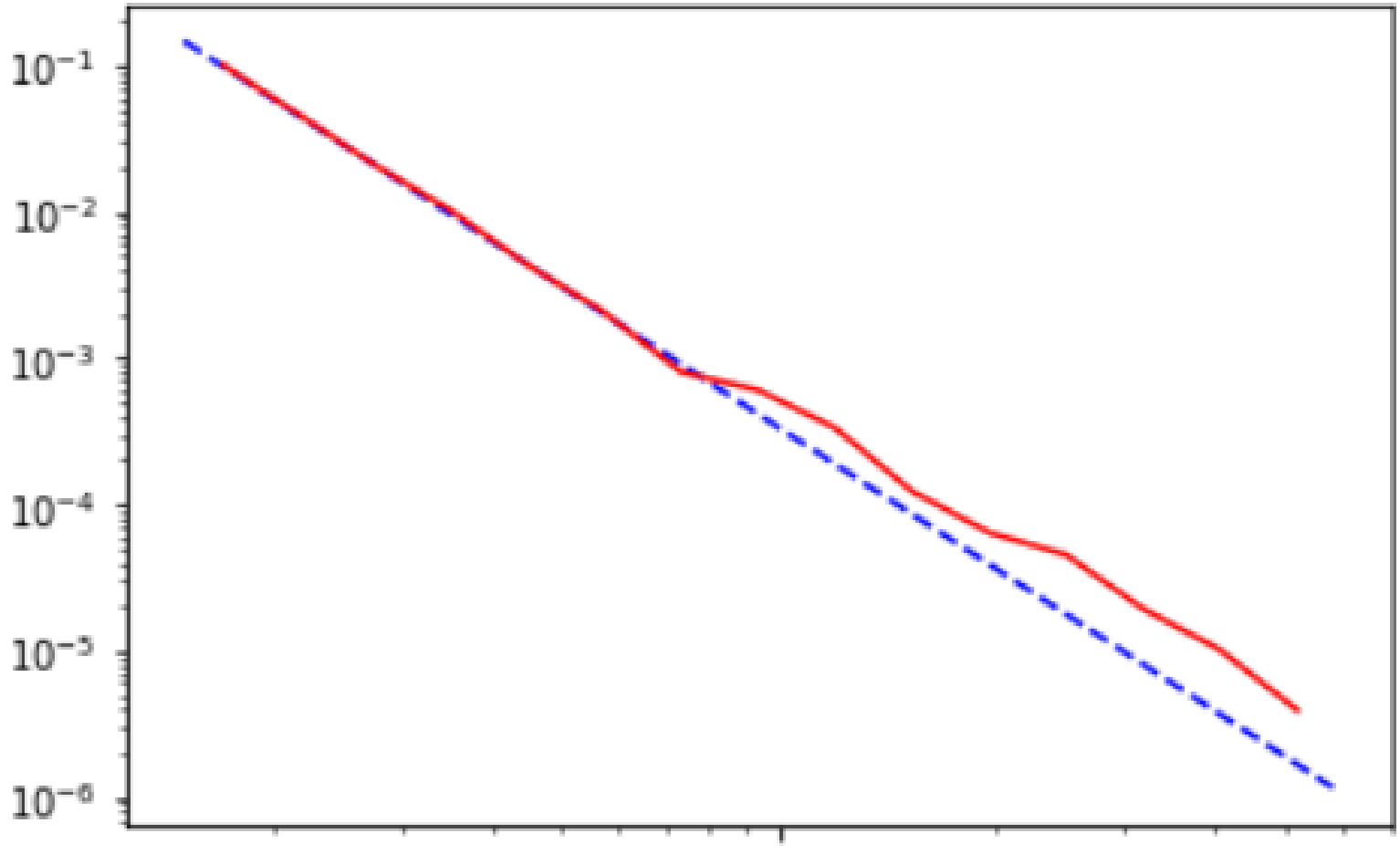
En posant $X=\log x$ *et* $Y=\log y$,

on trouve l'équation d'une fonction affine
$Y=\alpha X+\beta$
dont la pente α *est la valeur de l'exposant* k

Nombre de commentaires par commentateur

```
Calculating best minimal value for power law fit
xmin progress: 99%

<AxesSubplot:>
```



Modèle ERGM

- Les modèles ERGM (Exponential Random Graph Model) est un type de modélisation statistique qui vise à identifier, dans un graphe donné, quels sont les éléments qui déterminent et expliquent sa structure.
- Permettent d'identifier des motifs et d'évaluer les mécanismes sous-jacents à la formation des liens.
- L'objectif est d'estimer la probabilité que le réseau réel observé se forme en fonction de certaines de ses caractéristiques (locales ou globales) ou configurations spécifiques.
- On désigne par caractéristiques locales des motifs spécifiques, tels que la présence de triangles (trois nœuds connectés) ou de dyades
- On désigne par caractéristiques globales des statistiques résumant la structure du réseau, comme le nombre total de liens ou la densité du réseau.

Modèle ERGM: les étapes à suivre.

La démarche suppose plusieurs étapes :

- sélectionner les paramètres à entrer dans le modèle ;
- générer x centaines des graphes aléatoires (d'où le R de Random) de taille et de densité comparables au graphe étudié
- comparer statistiquement la distribution des paramètres dans le graphe étudié et dans les graphes aléatoires générés ;
- mesurer la qualité globale d'ajustement (goodness of fit) du modèle.

ERGM pondérés.

- On peut attribuer des poids aux différentes configurations de réseau plutôt qu'en considérant simplement leur présence ou leur absence. On parle alors de ERGM pondéré.

- Ainsi, les effets des motifs spécifiques peuvent être représentés soit de manière binaire (présence/absence) dans le cas des ERGM simples, soit de manière pondérée (avec des coefficients) dans le cas des ERGM pondérés. Le choix entre ces deux approches dépend de la complexité de la structure que l'on souhaite capturer dans le réseau étudié.

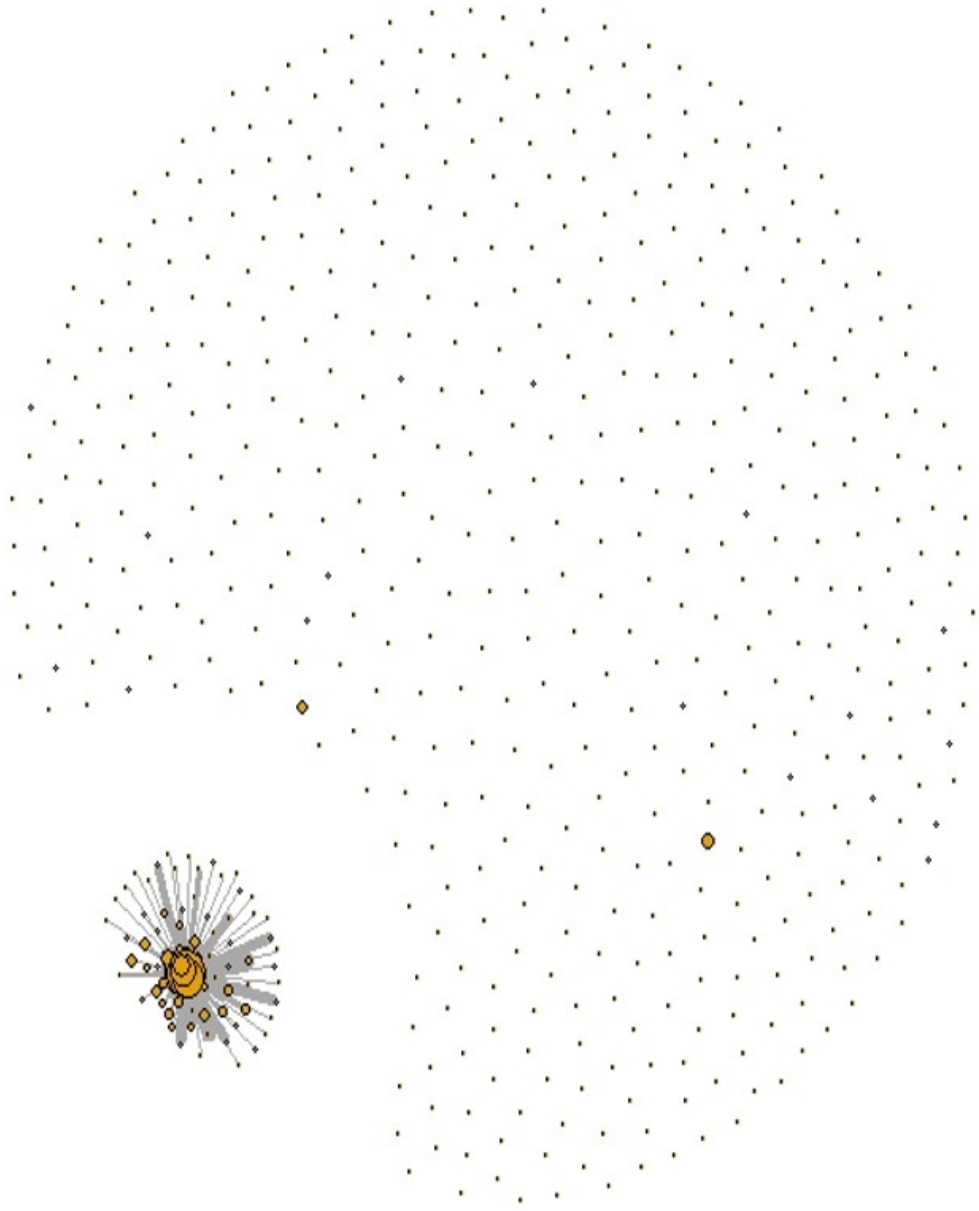
Réseau de commentateurs

On construit un réseau projeté (ou réseau homogène):

Dans ce réseau, il existe un lien entre deux commentateurs s'ils ont commenté au moins un channel en commun.

L'objectif est de comprendre si les commentaires se font de façon aléatoire (donneraient un graphe aléatoire) ou suivent une logique d'attachement.

Dans une première étape, on considère un ERGM simple, puis on projettera un ERGM pondéré.



ERGM pondéré

Graphe lourd à générer, difficultés techniques.
Décision prise de sous-échantillonner, selon les différentes stratégies des commentateurs.