Dynamiques des popularités dans YouTube

Cédric Richier* — **Georges Linares*** — **Rachid Elazouzi*** — **Tania Jimenez*** — **Eitan Altman**** — **Yonathan Portilla***

* *Laboratoire Informatique d'Avignon, 84000 Avignon, FRANCE*
** *INRIA, B.P 93, 06902 Sophia Antipollis Cedex, FRANCE*

RÉSUMÉ. *Cet article*[1] *est une étude de l'évolution du nombre de vues des contenus dans YouTube. Nous proposons dans un premier temps plusieurs modèles inspirés de l'économie et de la biologie pour caractériser les courbes d'évolution des nombres de vues des vidéos. Dans un deuxième temps, nous proposons une méthode automatique de classification de ces courbes en les associant à l'un des différents modèles suggérés. Nous montrons, sur un large ensemble de données, que* 90% *des vidéos peuvent être associées à l'un de ces modèles avec une erreur moyenne inférieure à* 5%*. Une étude empirique est menée au sujet de l'impact de la popularité et des catégories de vidéos sur l'évolution des nombres de vues. Enfin, cette classification est utilisée dans un exemple de méthode de prédiction de la popularité des vidéos.*

ABSTRACT. *The goal of this paper is to study the behaviour of view count in YouTube. We first propose several bio-inspired and economy-inspired models for the evolution of the view count of YouTube videos. We show, using a large set of empirical data, that the view count for 90% of videos in YouTube can indeed be associated to at least one of these models, with a Mean Error which does not exceed 5%. We derive automatic ways of classifying the view count curve into one of these models and of extracting the most suitable parameters of the model. We study empirically the impact of videos' popularity and category on the evolution of its view count. We finally use the above classification along with the automatic parameters extraction in order to predict the evolution of videos' view count.*

MOTS-CLÉS : *Média sociaux en ligne, popularité des vidéos, modèles de régression, prédiction de popularité*

KEYWORDS: *Online social media, video popularity, regression models, popularity prediction*

1. Ce travail est subventionné par la Commission Européenne dans le cadre du projet CONGAS FP7-ICT-2011-8-317672. Site web : www.congas-project.eu.

1. Introduction

Depuis sa création au début de l'année 2005, YouTube est devenu l'un des sites de partage de vidéos les plus populaire au monde. Aujourd'hui, il est le troisième site générant le plus de trafic sur Internet après Google et Facebook. L'un des aspects les plus important des vidéos de YouTube est leur popularité, décrite par leur nombre de vues. Comprendre et prédire cette popularité est d'un enjeu capital, notamment parce que les contenus plus visionnés génèrent du trafic et modéliser, estimer ou prédire la popularité a un impact direct sur les stratégies de mise en cache et de réplication que les diffuseurs devraient adopter. D'autre part, le modèle économique de l'Internet est basé sur la publicité et l'impact économique de la popularité est évident.

De nombreux chercheurs ont analysé les mécanismes de la popularité des vidéos générées par les utilisateurs (Cha *et al.*, 2007 ; Crane et Sornette, 2008 ; Gill *et al.*, 2007 ; Ratkiewicz *et al.*, 2010 ; Chatzopoulou *et al.*, 2010 ; Cha *et al.*, 2009). Si certains de ces travaux se sont concentrés sur la prédiction de la popularité (Szabo et Huberman, 2010), ses aspects temporels des dynamiques ont été relativement peu étudiés (Cha *et al.*, 2007 ; Cheng *et al.*, June, 2008 ; Mitra *et al.*, 2011).

Dans cet article, nous décrivons les comportements temporels les plus typiques du nombre de vues des vidéos dans YouTube et nous en proposons une modélisation qui permet de capturer les propriétés clés des dynamiques de popularité observées. Cette modélisation s'inspire de modèle utilisés notamment en biologie et repose sur l'hypothèse d'une forte similarité entre la propagation d'un contenu sur YouTube et celle d'une maladie infectieuse, thème classique de la biologie (Bailey, 1975 ; Meyers, 2007). De tels modèles ont déjà été utilisés , par exemple, pour décrire la propagation de virus dans les réseaux informatiques (Chakrabarti *et al.*, 2008 ; Ganesh *et al.*, 2005). Ils ont encore été considérés en marketing pour modéliser les dynamiques des cycles de vie de nouveaux produits (Bass, 1980). De nombreux articles en marketing ont montré que l'évolution des ventes des produits suivait un modèle de courbe en S où les ventes commencent à augmenter avec un taux croissant jusqu'à atteindre un régime stable à mesure que la limite du marché est approchée (Mahajan *et al.*, 2000).

La suite de l'article est structurée comme suit : dans la section 2 nous décrivons comment le jeu de données a été constitué. Nous proposons dans la section 3 les différents modèles dynamiques et leur utilisation. Notre méthode d'ajustement puis les résultats de la classification automatique et une application à la prédiction sont les propos des sections 4 et 5 respectivement. Enfin, la section 6 conclut l'article.

2. Collecte de données

Dans la mesure où nous voulons étudier différents types de dynamiques d'évolution du nombre de vues sur YouTube, il est nécessaire de disposer d'un grand nombre de vidéos. Dans cette section, nous décrivons comment notre jeu de données a été collecté. Sur YouTube, une vidéo est accompagnée d'un ensemble de méta données d'intérêt comme son titre, la date d'ajout au site, le nombre de vues, les vidéos recom-

mandées. La page web de la vidéo propose parfois quelques statistiques qui ne sont disponibles que si le propriétaire du contenu a donné son accord. Pour récupérer une partie de ces données, YouTube met à disposition deux interfaces de programmation (API) : l'API YouTube Data qui permet de récolter des données statiques, disponibles pour tout utilisateur (nombre de vues, titre, ...), et l'API YouTube Analytics qui permet de récupérer toutes sortes de statistiques comme les dynamiques de nombre de vues d'un contenu. Mais celle-ci n'est utilisable qu'avec autorisation du propriétaire du contenu. Par conséquent, une partie des données qui nous intéressent n'est pas accessible via les API. Nous utilisons ainsi un outil, nommé YOUStatAnalyzer (Zeni *et al.*, 2013), afin de récolter l'ensemble des données. Notre jeu de données contient plus de 80000 vidéos extraites de manière aléatoire dans YouTube et âgées entre 5 jours et 2500 jours. Pour chaque vidéo, nous avons retenu un ensemble d'informations statiques : l'identifiant YouTube, le titre, le nom de l'auteur, l'age et une liste de vidéos recommandées. Nous avons aussi stocké les évolutions de certaines métriques (partages, inscrits à la chaîne YouTube, temps de vue et nombre de vues) sous une forme journalière et cumulée journalière, depuis le jour d'ajout sur YouTube jusqu'au jour de la collecte.

3. Modèles de croissance de popularité

Nous concentrons notre analyse sur le nombre de vues comme métrique principale caractérisant la popularité des vidéos. Des travaux précédents ont montré une forte corrélation entre le nombre de vues et d'autres métriques comme le nombre de commentaires, l'ajout aux favoris et le score attribué par les utilisateurs. De plus, plus la popularité augmente, plus cette corrélation est forte (Chatzopoulou *et al.*, 2010). Nous supposons que les courbes de popularités en fonction du temps peuvent être catégorisées en un nombre réduit de types, modélisable par des classes de fonctions mathématiques relativement simples. Cette classification de nos modèles s'articule autour de deux critères :

– **Taille de la population potentielle :** Le premier critère est relatif à la taille de la population qui pourrait potentiellement être intéressée par le contenu. Nous différencions les modèles pour lesquels ce potentiel est considéré comme constant (nous parlerons alors de propriété de population potentielle fixe) de ceux dont ce potentiel est croissant dans le temps (nous parlerons de phénomène d'immigration, s'inspirant de la terminologie des processus de branchement).

– **Viralité :** Le second critère concerne la viralité structurelle d'une vidéo. Un modèle est considéré comme viral (ou ayant la propriété virale) si les utilisateurs ayant déjà visionné le contenu jouent un rôle significatif dans le processus de propagation (par le partage, par exemple). Il sera non viral si la propagation s'appuie essentiellement sur une diffusion massive depuis la source. Dans ce cas, une large partie de la population potentielle est informée de la présence du contenu directement par la source ou par des moyens de diffusion massifs (médias classiques, par exemple).

Dans la suite de cette section, nous décrivons les différents modèles dynamiques classés en fonction des deux critères précédents.

3.1. *Population potentielle fixe*

3.1.1. *Contenus viraux*

Pour caractériser les contenus viraux à population potentielle fixe, nous utilisons le *modèle Logistique* ou le *modèle de Gompertz*. Ces modèles sont, entre autre, utilisés pour la prévisions des ventes de produits de nouvelles technologies et font partie de la classe des modèles "en S", faisant référence à la forme de leurs courbes représentatives. Nous les avons considéré dans le cadre de la modélisation des évolutions du nombre de vues des vidéos car nous pensons qu'il peut y avoir une forte similarité entre une vidéo postée sur YouTube et le lancement d'un nouveau produit sur le marché des nouvelles technologies. En effet, comme il a été montré dans différents travaux en marketing, la vente de produits de nouvelles technologies se caractérise souvent par un début de croissance lente, suivi d'une croissance exponentielle rapide puis commence à ralentir pour venir s'écraser à l'approche de la limite du marché. Un tel comportement s'observe sur de nombreuses vidéos dans YouTube.

Modèle Logistique

Le *modèle Logistique* (aussi appelé *modèle Sigmoïde* dans cet article) est représenté par une fonction sigmoïde classique qui peut décrire l'évolution du nombre de vues de certaines vidéos à population potentielle fixe. Ce modèle est une équation différentielle non linéaire du premier ordre de la forme :

$$\frac{dS}{dt} = \lambda S(M - S) \qquad [1]$$

Où S représente le nombre de vues au temps t et M est la taille fixe de la population potentielle. C'est aussi une équation standard en épidémiologie pour décrire l'évolution du nombre d'individus infectés sous l'hypothèse d'une guérison impossible. Le nombre de nouveaux infectés à chaque instant est fonction d'un taux d'infection λ et de la taille de la population infectée S au temps t. Une solution de l'équation 1 est donnée par :

$$S(t) = \frac{M}{1 + (\frac{M - S(0)}{S(0)})e^{-\lambda M t}}$$

Cette fonction décrit une croissance exponentielle initiale suivie par une période dans laquelle le taux de croissance se réduit en s'approchant de la taille maximale de population potentielle. La courbe en S du *modèle Logistique* est symétrique. Cependant, dans le contexte des évolutions du nombre de vues, les phases convexes et concaves observées sont souvent non symétriques. Pour couvrir ces cas, nous considérons un autre modèle, le *modèle de Gompertz*.

Modèle de Gompertz

Un modèle qui répond au problème posé par l'aspect symétrique du *modèle Logistique* est donné par l'équation dynamique suivante :

$$\frac{dS}{dt} = \lambda S \log(\frac{M}{S}) \quad [2]$$

Ce modèle est appelé *modèle de Gompertz*. Il est utilisé comme modèle de croissance de certaines tumeurs cancéreuses et aussi comme modèle de diffusion de produits dans les marchés. Une solution de l'équation 2 est donnée par la fonction de Gompertz :

$$S(t) = M \exp\left(-\log(\frac{M}{S(0)}) \exp(-\lambda t)\right)$$

Ce modèle est similaire au *modèle Logistique* mais permet la non-symétrie par rapport au point d'inflexion. En général, le *modèle de Gompertz* atteint l'inflexion plut tôt dans le processus de croissance. Ce comportement semble bien convenir pour décrire un grand nombre de dynamiques de nombre de vues observées sur YouTube.

3.1.2. *Contenus non viraux*

Un contenu non viral est la résultante d'une situation où les utilisateurs ne contribuent pas de manière significative à sa propagation. Cela peut être le cas de contenus qui gagnent de la popularité à travers des campagnes de publicité ou d'autres outils de marketing. Par exemple, une publicité sur l'existence du contenu peut être diffusée de manière large sur un réseau social et ensuite, les utilisateurs ainsi informés accèdent au contenu de façon aléatoire. Pour ces situations, nous suggérons un modèle dynamique suivant l'équation différentielle linéaire :

$$\frac{dS}{dt} = \lambda(M - S) \quad [3]$$

Ce modèle est appelé *modèle exponentiel négatif*. La solution de 3 est donnée par :

$$S(t) = S(0) + (M - S(0))(1 - e^{-\lambda t})$$

3.2. *Phénomène d'immigration*

L'hypothèse d'une population potentielle fixe est souvent une approximation raisonnable lorsque l'évolution de la popularité d'un contenu augmente rapidement puis meurt dans un intervalle de temps assez court. Mais, pour de nombreux cas, cette hypothèse devient inappropriée lorsque le temps mis pour atteindre le régime saturé est plus long. Ici, nous considérons le cas du phénomène d'immigration où la croissance de la population potentielle et la dynamique du nombre de vues sont liées de manière complexe. Pour prendre en compte cette dépendance, nous envisageons différents scénarios de croissance qui modélisent les cas viraux et non viraux. Dans cet article, nous restreignons notre étude au cas où la population potentielle augmente à vitesse constante.

3.2.1. *Contenus viraux*

Nous nous intéressons ici au phénomène d'immigration dans le cas des contenus viraux. Dans ce type de dynamique, la courbe d'évolution du nombre de vues adopte d'abord un comportement viral (dans une forme en S) puis fini par croître linéairement. Une solution possible pour décrire un tel comportement consiste à ajouter une composante linéaire à la fonction de Gompertz :

$$S(t) = M \exp\left(-\log(\frac{M}{S(0)})\exp(-\lambda t)\right) + kt$$

Cette dynamique, que nous nommons *modèle de Gompertz modifié*, semble convenir à bon nombre d'exemples présents dans notre jeu de données.

3.2.2. *Contenus non viraux*

Le *modèle de croissance linéaire* $S(t) = S(0)+\lambda t$ décrit simplement une situation où les utilisateurs ne propagent pas le contenu aux autres mais où la vidéo bénéficie d'un phénomène d'immigration dont résulte une croissance linéaire du nombre de vues.

Un autre type de courbes non virales observées sont des courbes concaves (données par le *modèle exponentiel négatif*) qui ne convergent pas vers une asymptote horizontale mais deviennent linéaires à l'horizon, sous l'influence du phénomène d'immigration. De telles dynamiques peuvent se modéliser en modifiant les solutions de l'équation 3 par l'ajout d'une composante linéaire. Cela nous donne le *modèle exponentiel négatif modifié* :

$$S(t) = S(0) + (M - S(0))(1 - e^{-\lambda t}) + kt$$

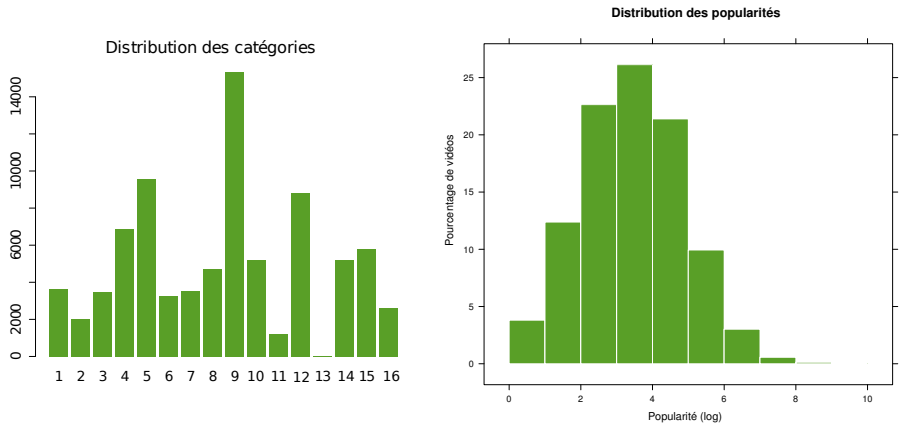
où k est le taux de croissance de la population potentielle. Dans ce cas, M représente alors la taille de population à partir de laquelle le processus devient linéaire.

4. Méthode d'ajustement

4.1. *Jeu de données*

Comme nous l'avons expliqué dans la section 2, nous avons collecté des méta informations sur plus de 80000 vidéos. En plus de la dynamique des nombres de vues que nous utilisons pour l'ajustement des modèles, les différentes caractéristiques que nous considérons dans cette étude pour chaque vidéo sont leur age (en nombre de jours), la catégorie YouTube et la popularité (nombre total de vues au jour de la collecte). La Figure 1b montre la distribution de la popularité des vidéos dans le jeu de données selon une échelle logarithmique.

La Table 1 liste les catégories YouTube contenues dans le jeu de données et la Figure 1a en illustre leur distribution. Un résumé des ages et des valeurs du nombre de vues total est présenté dans la Table 2.



(a) Distribution des catégories (b) Distribution logarithmique des popularités

Figure 1. *Distributions de quelques caractéristiques présente dans le jeu de données*

Tableau 1. *Liste des catégories représentées dans le jeu de données*

1. "Animals"	7. "Games"	13. "Shows"
2. "Autos"	8. "Howto"	14. "Sports"
3. "Comedy"	9. "Music"	15. "Tech"
4. "Education"	10. "News"	16. "Travel"
5. "Entertainment"	11. "Nonprofit"	
6. "Film"	12. "People"	

Tableau 2. *Résumé des ages et des popularités dans le jeu de données YouTube*

Age (jours)	**Popularité (nombre de vues)**
Minimum : 5	Minimum : 1
1^{er} Quartile : 140	1^{er} Quartile : $2,650.10^2$
Médiane : 393	Médiane : $2,728.10^3$
Moyenne : $610,5$	Moyenne : $6,091.10^5$
3^{ime} Quartile : 923	3^{ime} Quartile : $2,630.10^4$
Maximum : 2426	Maximum : $1,746.10^9$

4.2. *Ajustement*

4.2.1. *Observations et normalisation*

Pour l'ajustement des modèles aux données, nous n'utilisons que l'évolution cumulée du nombre de vues en fonction du temps. Nous désignons un ensemble d'obser-

vations d'une vidéo par : $(Y_i, i)_{1 \leq i \leq n}$ où Y_i est le nombre cumulé de vues au jour i et n est le nombre total d'observations (l'age de la vidéo en nombre de jours). Afin d'éviter quelques problèmes techniques lors de l'utilisation de l'algorithme d'estimation, nous utilisons des observations normalisées : $(y_i = \frac{Y_i}{Y_n}, t_i = \frac{i}{n})_{1 \leq i \leq n}$.

4.2.2. *Méthode d'estimation des paramètres*

Nous estimons les paramètres des modèles décrit dans la section 3 en utilisant un algorithme de régression basé sur la minimisation du critère des moindres carrés. Étant donné un ensemble d'observations normalisées $(y_i, t_i)_{1 \leq i \leq n}$, soit S l'expression d'un des modèles considérés. Le critère des moindres carrés (MSC, Mean Squares Criterion en anglais) est défini par : $MSC = \sum_i (S(t_i) - y_i)^2$. La méthode que nous utilisons est l'algorithme de Levenberg-Marquardt (Marquardt, 1963) qui est connu pour son efficacité dans les cas non linéaires. C'est un processus itératif d'estimation des paramètres des modèles par la résolution d'un problème de minimisation du MSC. Une formulation explicite des modèles doit être connue pour cette méthode car le calcul des dérivées partielles est nécessaire tout au long du processus. L'un des défauts de cet algorithme, comme toute autre méthode non linéaire, est que la solution obtenue n'est pas nécessairement globale. Malgré cela, cet algorithme nous a permis d'obtenir de très bons résultats lors de nos expériences.

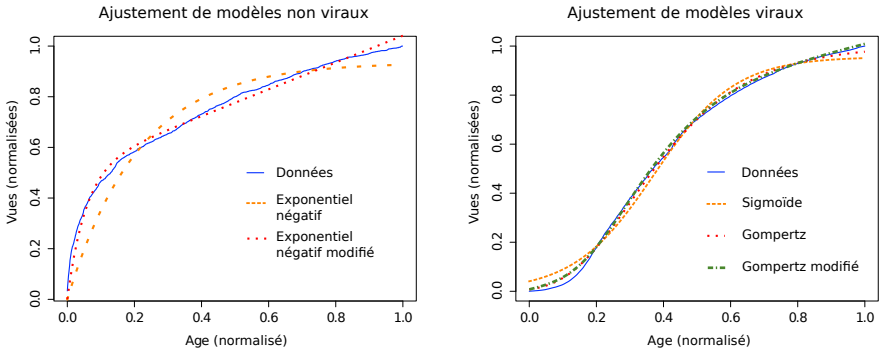
4.2.3. *Ajustement aux données pour les contenus non viraux*

La dynamique donnée par l'équation 3 convient pour les contenus dont la courbe du nombre de vues est concave puis s'aplatit en approchant la limite de population potentielle. Dans la Figure 2a, nous donnons un exemple d'application de ce modèle. Nous observons que la courbe estimée (en tirets) admet une asymptote de pente nulle. Mais, la courbe représentant les vraies données (ligne pleine) semble suivre une ligne oblique à l'horizon. Cela nous indique que la population potentielle croît probablement en fonction du temps et que nous sommes donc face à un phénomène d'immigration. Dans ce cas, nous modélisons la dynamique avec le *modèle exponentiel négatif modifié* introduit dans la sous-section 3.2.2. La courbe ajustée de ce modèle (en pointillés) semble mieux convenir.

4.2.4. *Ajustement aux données pour les contenus viraux*

Trois modèles sont considérés pour le cas des contenus viraux : le *modèle Logistique* et le *modèle de Gompertz* dans le cadre d'une population potentielle fixe et le *modèle de Gompertz modifié* dans le cadre d'une population potentielle croissante (voir 3.2.1). La

. 2b est un exemple d'ajustement de ces modèles à un contenu YouTube. Nous observons que la forme en S du *modèle Logistique* est symétrique (tirets). Cependant, les phases convexe et concave ne le sont pas, comme le montre la courbe des données (pleine). Le *modèle Logistique* n'est donc pas adéquat. Ensuite, le *modèle de Gompertz* et le *modèle de Gompertz modifié* sont ajustés au même contenu. Le *modèle de Gompertz* (pointillés) convient mieux que le *modèle Logistique*, et le *modèle de Gom-*



(a) Contenu non viral (b) Contenu viral

Figure 2. *Exemples d'ajustement de différents modèles à un contenu non viral (2a) et viral (2b)*

pertz modifié (points-tirets) décrit mieux le comportement à l'horizon (phénomène d'immigration).

5. Classification automatique

Une contribution importante de notre travail consiste à proposer un système de classification automatique des contenus YouTube en leur associant l'un des modèles dynamiques. Pour chaque contenu, nous devons faire face à deux problématiques : d'abord, chacun des modèles doit être évalué afin de sélectionner des candidats puis les modèles retenus doivent être comparés afin d'en choisir le meilleur. Considérons d'abord la question de l'évaluation de chaque modèle. Comme expliqué dans la section 4.2.2, l'estimation des paramètres est basée sur la minimisation du critère des moindres carrés. Nous définissons le taux d'erreur moyen (MER, Minimum Error Rate en anglais) par :

$$MER = \frac{1}{n} \sum_i \frac{|S(t_i) - y_i|}{y_i + 1}$$

Le MER représente le taux d'erreur moyen fait par un modèle S au regard des observations. Par exemple, si $MER \leq 0,05$, on peut dire qu'en moyenne, l'erreur de l'estimation à chaque instant est inférieure à 5% comparée à la valeur réelle du nombre de vues. Le choix du terme (y_i+1) au dénominateur vise à lisser les grandes valeurs du MER pouvant être générées par quelques très petites valeurs des y_i. Cette métrique nous permet de fixer un seuil au delà duquel un modèle peut être considéré comme non satisfaisant. Pour comparer les modèles satisfaisants, nous introduisons un critère de qualité discuté dans (Deming, 1934). Pour formuler ce critère, nous définissons d'abord le degré de liberté d'un modèle par $df = n - p$ où p est le nombre de para-

Tableau 3. *GoF pour les modèles de la Figure 2a*

Modèle	MSC	MER	GoF
Exponentiel négatif	3.558	0.074	0.004
Exponentiel négatif modifié	0.453	0.027	$4.98.10^{-4}$

Tableau 4. *GoF pour les modèles de la Figure 2b*

Modèle	MSC	MER	GoF
Sigmoïde	0.480	0.021	10^{-3}
Gompertz	0.092	0.018	$1.846.10^{-4}$
Gompertz modifié	0.033	0.008	$8.831.10^{-5}$

mètres du modèle et n le nombre de données. Le critère de qualité (GoF, goodness of fit en anglais) est alors donné par :

$$GoF = \frac{1}{(df)} MSC$$

Le modèle présentant le plus petit GoF sera alors considéré comme le meilleur. Dans les Tables 3 et 4, nous donnons les valeurs de MSC, MER et GoF pour les modèles utilisés dans la Figure 2a et la Figure 2b respectivement. Dans l'exemple de la Figure 2a, avec un seuil du MER fixé à 0.075, le *modèle exponentiel négatif* et le *modèle exponentiel négatif modifié* sont tous deux pertinents. Avec une valeur du GoF à $4.98.10^{-4}$, le *modèle exponentiel modifié* est le plus adapté. Dans l'exemple donné dans la Figure 2b, si le seuil du MER est fixé à 0.02, le *modèle Sigmoïde* (i.e le *modèle Logistique*) n'est pas pertinent alors que le *modèle de Gompertz* et sa version modifiée respectent tous deux la contrainte du seuil. Si l'on se fie à la valeur du GoF, le *modèle de Gompertz modifié* est le meilleur avec $GoF = 8.831.10^{-5}$. Pour aller plus loin, la problématique du seuil pour le MER est cruciale afin de s'appuyer sur un filtre acceptable pour de nombreuses vidéos.

La prochaine étape consiste à associer à chaque vidéo du jeu de donnée un modèle mathématique en se basant sur le MER et le GoF. Notre méthode de classification montre que presque 90% des vidéos peuvent être associées à l'un des six modèles avec un MER inférieur à 0.05 (voir Table 5). Un taux d'erreur moyen de 5% nous semble raisonnable pour dire que l'ajustement est fiable. Remarquons que si le seuil du MER

Tableau 5. *Pourcentage des contenus par tranches de valeur du MER*

MER	$\leq 5\%$	$]5\%; 10\%]$	$> 10\%$
Vidéos	$\mathbf{88.79\%}$	8.51%	2.70%

est fixé à 0.1, plus de 97% des vidéos correspondent à l'un des modèles. Il y a donc moins de 3% des vidéos pour lesquelles nos modèles donnent un taux d'erreur élevé

(plus de 10%). La Figure 3 illustre un exemple d'une telle vidéo. L'association n'est pas fiable essentiellement à cause des différents changements de comportement. De ce fait, il semble que nos modèles ne soient pas fiables pour capturer l'effet de multiples pics dans l'évolution du nombre de vues. L'étude de tels cas fera l'objet de prochains travaux. Dans la Figure 4a, nous montrons la distribution du MER pour chaque mo-

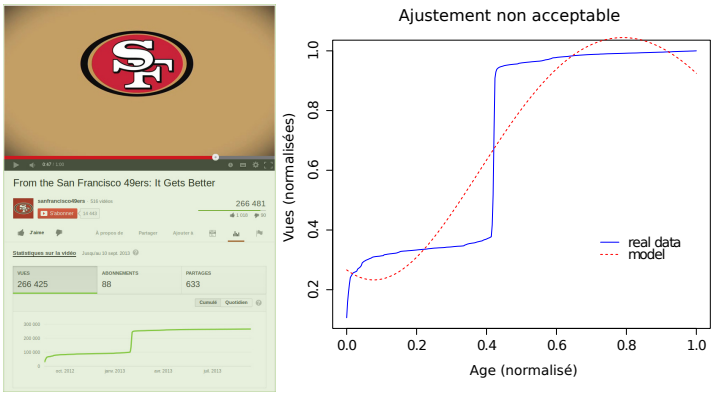
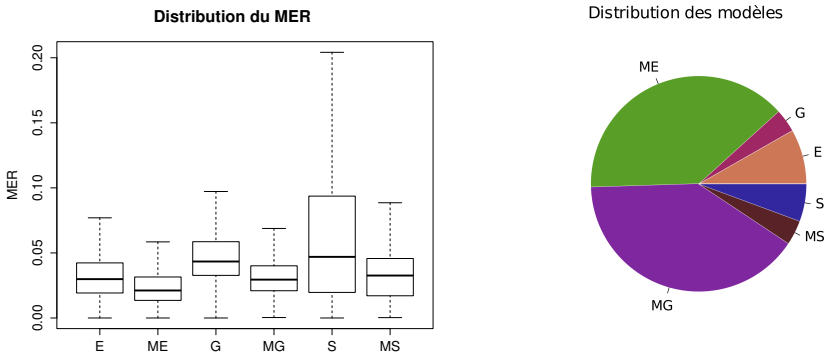


Figure 3. *Exemple d'ajustement non acceptable*

dèle. Nous considérons aussi un nouveau modèle, le *modèle sigmoïde modifié*, dérivé du *modèle Logistique* par l'ajout d'une composante linéaire (comme décrit dans la section 3.2.1 pour le *modèle de Gompertz modifié*). Dans cette figure, E, ME, G, MG, S et MS font respectivement référence aux modèles *exponentiel négatif*, *exponentiel négatif modifié*, *Gompertz*, *Gompertz modifié*, *sigmoïde* et *sigmoïde modifié*. Il apparaît que les modèles *ME* et *MG* donnent des ajustement avec moins d'erreur que les autres. On peut les considérer comme les modèles les plus fiables pour notre jeu de données. De plus, ces deux modèles couvrent plus de 75% des vidéos (voir Fig. 4b). Les deux modèles entrent dans le cadre du phénomène d'immigration. Étant donnés ces résultats, nous pouvons conclure qu'une large part des vidéos dans YouTube continuent à susciter de l'intérêt après une longue période. De plus, il apparaît un certain équilibre entre les contenus viraux et non viraux.

Il est maintenant naturel de se demander si la distribution donnée par la classification reste la même selon les catégories YouTube ou la popularité finale d'une vidéo. Pour répondre à ces questions, nous avons fait la classification dans quatre catégories principales suggérées par la distribution donnée dans la Figure 1a : Music (plus de 14000 vidéos), Entertainment(plus de 8500 vidéos), People (autour de 7500 vidéos) et Education (presque 6000 vidéos). En général, la distribution des modèles dans chacune de ces catégories est semblable et montre peu de dépendance aux thèmes des vidéos, à l'exception de la catégorie Education où plus de 50% des vidéos suivent un *modèle de Gompertz modifié*. De plus, les modèles viraux couvrent presque 75%



(a) Distribution du MER par modèle (b) Distribution des modèles après classification

Figure 4. *Analyses de la classification automatique des processus de diffusion dans YouTube*

des vidéos. Dans cette catégorie, il semble que le bouche à oreille soit le mécanisme dominant suivant lequel les contenus se propagent.

Nous analysons aussi la distribution des modèles en considérant différentes tranches de popularité. Selon la distribution illustrée dans la Figure 1b, nous définissons sept classes de popularité listées dans la Table 6. La distribution des modèles

Tableau 6. *Classes de popularité*

Classes de popularité	**Nombre total de vues** V
Extremely unpopular (EUP)	$0 \leqslant V < 10$
Very unpopular (VUP)	$10 \leqslant V < 100$
Unpopular (UP)	$100 \leqslant V < 1000$
Not so popular (NSP)	$1000 \leqslant V < 10^4$
Popular (P)	$10^4 \leqslant V < 10^5$
Very popular (VP)	$10^5 \leqslant V < 10^6$
Extremely popular (EP)	$10^6 \leqslant V$

pour chaque classe de popularité est donnée dans la Table 7.

Nous pouvons observer que les distributions varient selon les classes de popularité. D'abord, le *modèle Sigmoïde* est dominant pour les vidéos extrêmement impopulaires (comptabilisant moins de 10 vues). Les vidéos populaires et moyennement populaires peuvent être groupées en terme de distribution des modèles avec près de 50% pour le *modèle de Gompertz modifié* et 35% pour le *modèle exponentiel modifié*. Les vidéos très populaires et extrêmement populaires peuvent aussi être groupées avec une distribution proche de celle obtenue sur l'ensemble du jeu de données (voir Figure 4b). Les

Tableau 7. *Distribution des modèles par classes de popularité (en %)*

Model	EUP	VUP	UP	NSP	P	VP	EP
E	11.4	12.2	8.4	8	6.8	6.2	5.7
G	1.6	2.8	1.8	2.5	3.6	3.2	2.1
ME	11.6	54.5	48.9	35.2	35.1	42.3	47.5
MG	2.5	19.4	34.7	48.7	49.3	44.3	42.8
MS	1.8	4.3	4.3	3.6	2.9	2.3	0.8
S	70.7	6.6	1.5	1.7	2	1.3	0.8

vidéos impopulaires et très impopulaires suivent un *modèle exponentiel modifié* dans 50% des cas. Le *modèle de Gompertz modifié* représente moins de 20% pour les vidéos très impopulaires alors qu'il couvre presque 35% des vidéos impopulaires. Dans la suite, nous introduisons brièvement une méthode de prédiction qui s'appuie sur la classification des vidéos.

Méthode de prédiction

Dans cette section, nous présentons une méthode de prédiction de l'évolution du nombre de vues à partir d'une date donnée t_f jusqu'à une date cible t_p. Nous appelons fenêtre de prédiction T la différence entre t_p et t_f. Cette prédiction se base sur l'historique connue des vidéos donnée par un ensemble d'observations $(y_i, t_i)_{1 \leq i \leq f}$ jusqu'au temps t_f où f est le nombre d'observations[1]. En combinant ces informations avec les modèles de la classification, l'évolution du nombre de vues est estimée en sélectionnant l'un des modèles ajustés aux observations. En appliquant la méthode à notre jeu de données, nous évaluons une taille maximale de la fenêtre de prédiction confinée à moins de 5% d'erreur moyenne. Plus formellement :

$$T_{max} = \max\{t_p - t_f | \frac{1}{p-f} \sum_{i=f+1}^{p} \frac{|S_{t_f}(t_i) - y_i|}{(y_i + 1)} \leq 0.05\}$$

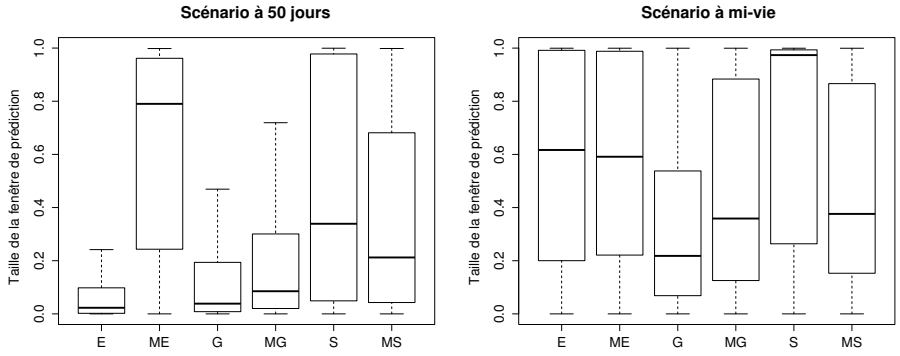
où S_{t_f} est le modèle mathématique sélectionné.

Nous avons testé notre méthode de prédiction dans un scénario où t_f correspond à la moitié de la vie des vidéos dans le jeu de données. Soit $\Delta T = \frac{T_{max}}{t_n - t_f}$ où $t_n - t_f$ est le temps restant de la vidéo depuis t_f. Notons que ΔT est borné par 1. La Figure 5b donne une vision de la dispersion de ΔT en fonction des modèles sélectionnés. La Table 8 précise les valeur de la moyenne et de la variance de ΔT pour chaque modèle ainsi que le nombre de vidéos classifiées dans les différents modèles. Nos résultats soulignent une capacité de prédiction intéressante et la plupart des modèles donnent des fenêtres de prédiction, contraintes à 5% d'erreur moyenne, suffisamment longues.

1. Le jeu de données utilisé pour la prédiction contient des vidéos âgées d'au moins 50 jours.

Tableau 8. *Moyenne et variance, par modèle, de la taille de la fenêtre de prédiction dans le scénario mi-vie*

Modèle	moyenne	variance	nombre de vidéos
E	0.58	0.15	4132
ME	0.58	0.14	25281
G	0.34	0.12	683
MG	0.46	0.14	19349
S	0.68	0.15	1030
MS	0.46	0.13	1659



(a) Scénario 50 jours

(b) Scénario mi-vie

Figure 5. *Taille de la fenêtre de prédiction par modèle*

Plus précisément, avec un ΔT moyen autour de 0.5, en ayant observé la moitié de la vie d'une vidéo, notre méthode permet en moyenne de prédire l'évolution future du nombre de vues jusqu'à la moitié du temps restant.

Nous avons fait le choix ici de baser la prédiction sur une séquence d'observation longue de la moitié de la vie des vidéos dans le jeu de données. Cela permet de s'appuyer sur la même quantité de données relative, indépendamment de la durée effective de la vie d'une vidéo. Nous comparons ensuite ces résultats à un scénario où la longueur de la séquence observée est fixée à 50 jours. Notons que 50 jours représentent bien moins que la moitié de la vie de la plupart des vidéos dans le jeu de données et que donc la prédiction est moins efficace. Les résultats correspondant à ce nouveau scénario sont donnés dans la Table 9 et la Figure. 5b. Nous observons cependant des résultats similaires de la taille moyenne des fenêtres de prédiction pour les modèles *exponentiel modifié* et *Sigmoïde* (i.e *Logistique*).

Tableau 9. *Moyenne et variance de la taille de la fenêtre de prédiction pour le scénario à* 50 *jours*

Modèle	moyenne	variance	nombre de vidéos
E	0.12	0.06	3401
ME	0.62	0.14	21788
G	0.19	0.09	687
MG	0.23	0.09	13821
S	0.48	0.11	1137
MS	0.21	0.18	1561

6. Conclusions et perspectives

Dans ce travail, nous avons proposé une méthode automatique de classification des dynamiques des nombres de vues des vidéos selon six modèles issus de la biologie ou de l'économie. Nous avons montré la bonne couverture de l'ensemble d'échantillons avec ces six modèles et la correspondance quasi parfaite entre les modèles et les données observées. Cette approche permet de caractériser les dynamiques des popularités selon deux propriétés que sont la viralité et la croissance de la population potentielle. Ensuite, deux des six modèles apparaissent souvent comme les meilleurs candidats pour l'ajustement aux données et ces deux modèles soulignent le fait que la plupart des vidéos bénéficient d'un phénomène d'immigration résultant en une attraction continue des utilisateurs à long terme. Finalement, nous proposons une méthode de prédiction de l'évolution du nombre de vues sur une fenêtre de temps qui donne des résultats encourageants.

Sur la base de ce travail, nous identifions plusieurs directions pour de futurs travaux. (i) Collecter une plus grande base de données dans la mesure où certaines caractéristiques ne pourraient être détectées que sur un très large échantillon. (ii) Affiner les résultats obtenus sur la distribution des modèles en prenant en considération des caractéristiques liées aux propriétaires de contenus (par exemple le réseau des diffuseurs de vidéos, les inscrits à une chaîne, l'audience des vidéos précédemment postées sur une chaîne, etc) ce qui pourrait permettre d'améliorer la prédiction de la popularité future.

7. Bibliographie

Bailey N., *The Mathematical Theory of Infectious Diseases and its Applications*, Griffin, London, 1975.

Bass F. M., « The Relationship Between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durable Technological Innovations », *The Journal of Business*, vol. 53, n° 3, p. pp. 51-67, 1980.

Cha M., Kwak H., Rodriguez P., Ahn Y.-Y., Moon S., « I tube, you tube, everybody tubes : analyzing the world's largest user generated content video system », *Proc. of ACM IMC*, San Diego, California, USA, p. 1-14, October 24-26, 2007.

Cha M., Kwak H., Rodriguez P., Ahn Y.-Y., Moon S., « Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems », *IEEE/ACM Transactions on Networking*, vol. 17, n° 5, p. 1357 - 1370, 2009.

Chakrabarti D., Wang Y., Wang C., Leskovec J., Faloutsos C., « Epidemic Thresholds in Real Networks », *ACM Trans. Inf. Syst. Secur.*, vol. 10, n° 4, p. 1 :1-1 :26, jan, 2008.

Chatzopoulou G., Sheng C., Faloutsos M., « A First Step Towards Understanding Popularity in YouTube », *in Proc. of IEEE INFOCOM*, San Diego, p. 1 -6, March 15-19, 2010.

Cheng X., Dale C., Lui J., « Statistics and Social Network of YouTube Videos », *In Proc. International Workshop on Quality of Service (IWQoS) The Netherlands*, vol. , p. 229 – 238, June, 2008.

Crane R., Sornette D., « Viral, quality, and junk videos on YouTube : Separating content from noise in an information-rich environment », *Proc. of AAAI symposium on Social Information Processing*, Menlo Park, California, CA, March 26-28, 2008.

Deming W. E., « The Chi-Test and Curve Fitting », *Journal of the American Statistical Association*, vol. 29, n° 188, p. 372-382, Dec, 1934.

Ganesh A., Massoulie L., Towsley D., « The effect of network topology on the spread of epidemics », *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 2, Miami, FL, USA, p. 1455-1466, March, 2005.

Gill P., Arlitt M., Li Z., Mahanti A., « YouTube Traffic Characterization : A View From the Edge », *Proc. of ACM IMC*, 2007.

Mahajan V., Muller E., Wind Y., *New-Product Diffusion Models*, International Series in Quantitative Marketing, Springer, 2000.

Marquardt D. W., « An Algorithm for Mean-Squares Estimation of Nonlinear Parameters », *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, n° 2, p. 431-441, jun, 1963.

Meyers L. A., « Contact network epidemiology : Bond percolation applied to infectious disease prediction and control », *Bull. AMS*, vol. 44, n° 1, p. 63-86, 2007.

Mitra S., Agrawal M., Yadav A., Carlsson N., Eager D., Mahanti A., « Characterizing Web-based Video Sharing Workloads », *ACM Transactions on the Web*, vol. 2, n° 8, p. 8 - 27, 2011.

Ratkiewicz J., Menczer F., Fortunato S., Flammini A., Vespignani A., « Traffic in Social Media II : Modeling Bursty popularity », *Proc. of IEEE SocialCom*, Minneapolis, August 20-22, 2010.

Szabo G., Huberman B. A., « Predicting the Popularity of Online Content », *Communications of the ACM*, vol. 53, n° 8, p. 80-88, aug, 2010.

Zeni M., Miorandi D., De Pellegrini F., « YOUStatAnalyzer : a Tool for Analysing the Dynamics of YouTube Content Popularity », *Proc. 7th International Conference on Performance Evaluation Methodologies and Tools (Valuetools, Torino, Italy, December 2013)*, Torino, Italy, 2013.