Automatic role detection in online forums

Alberto Lumbreras

Détection automatique des rôles dans les forums en ligne

Alberto Lumbreras
Université Lumière Lyon 2

October 26, 2016

Abstract

Nous traitons dans cette thèse le problème de la détection des rôles des utilisateurs sur des forums de discussion en ligne. On peut définir un rôle comme l'ensemble des comportements propres d'une personne ou d'une position. Sur les forums de discussion, les comportements sont surtout observés à travers des conversations. Pour autant, nous centrons notre attention sur la manière dont les utilisateurs dialoguent. Nous proposons trois méthodes pour détecter des groupes d'utilisateurs où les utilisateurs d'un même groupe dialoguent de façon similaire.

Notre première méthode se base sur les structures des conversations dans lesquelles les utilisateurs participent. Nous appliquons des notions de voisinage différentes (*radius-based, order-based* et *time-based*) applicables aux commentaires qui sont représentés par des nœuds sur un arbre. Nous comparons les motifs de conversation qu'ils permettent de détecter, ainsi que les groupes d'utilisateurs associés à des motifs similaires.

Notre deuxième méthode se base sur des modèles stochastiques de croissance appliqués aux fils de discussion. Nous proposons une méthode pour trouver des groupes d'utilisateurs qui ont une tendance à répondre au même type de commentaire. Nous montrons que, bien qu'il y ait des groupes d'utilisateurs avec des motifs de réponse similaires, il n'y a pas d'évidence forte qui confirme que ces comportements présentent des propriétés prédictives quant aux comportements futurs –sauf pour quelques groupes avec des comportements extrêmes.

Avec notre troisième méthode nous intégrons les types de données utilisés dans les deux méthodes précédentes (*feature-based* et *behavioral* ou *functional-based*) et nous montrons que le modèle trouve des groupes en ayant besoin de moins d'observations. L'hypothèse du modèle est que les utilisateurs qui ont des caractéristiques similaires ont aussi des comportements similaires.

1 Introduction

1.1 Contexte et motivation

Dès les premiers newsgroups dans les années 1980 jusqu'à nos jours, les forums en ligne ont toujours été parmi les moyens les plus populaires de communication en ligne. Même après l'énorme expansion des réseaux sociaux, les forums sont encore utilisés par le 15% des utilisateurs en ligne aux US (Duggan, 2015). Les forums modernes couvrent une grande variété de sujets tels que la politique, la

santé, la technologie ou les jeux vidéo, et aussi de nombreuses applications telles que les questions& réponses (Q & A), le partage et la discussion de nouvelles (newsboards), la recherche d'aide des autres étudiants sur les *Massive Open Online Courses* (MOOC), ou la discussion entre partisans d'une même cause politique. La croissance actuelle des forums comme Quora, Reddit ou StackExchange suggère que cette forme de communication en ligne est plus forte que jamais et qu'il a encore un énorme potentiel en termes de nombre d'utilisateurs et des nouvelles applications.

Alors que les forums deviennent de plus en plus peuplés et les utilisateurs produisent de plus en plus de contenu, ils ouvrent la porte à de nouveaux défis et opportunités dans des domaines comme l'informatique, les systèmes complexes et la sociologie. Les informaticiens peuvent développer des outils pour aider les utilisateurs à explorer le contenu des forums afin que l'expérience soit satisfaisante et les utilisateurs ne pas abandonner la communauté ; des domaines tels que l'apprentissage automatique, la recherche d'information et les systèmes de recommandation commencent à jouer un rôle important ici. Le cadre des réseaux complexes est excellent pour analyser et comprendre comment les nouvelles dynamiques émergent du niveaux *micro* au *macro*, ou de l'individu à la communauté. Certains modèles ont été proposés, par exemple, pour expliquer comment la structure d'une conversation se développe au fil du temps, et il est étonnant de voir comment certains modèles mathématiques de la dynamique humaine ont la simplicité des lois physiques (Kumar et al., 2010; Gómez et al., 2012).

Les forums en ligne sont aussi des lieux où des communautés émergent (Kollock, 1998). Les réseaux des interactions entre les utilisateurs sous la forme de messages ou des commentaires finit par créer des significations partagées et, en général, une culture commune qui définit et délimite l'ensemble des comportements possibles. A titre d'exemple, le slogan populaire *don't feed the troll* suggère aux utilisateurs d'ignorer ceux qui jouent le rôle d'un *troll*.

Néanmoins, les approches les plus fructueuses ont besoin d'être interdisciplinaires. Les forums sont associés à un grand volume de données, à des systèmes complexes et des communautés humaines et, en tant que tels, une vision intégrée est susceptible d'apporter des résultats plus fructueux que la somme des parties (McFarland et al., 2016; Tinati et al., 2014).

Le thème central de cette thèse est l'analyse des *rôles en ligne*. Les rôles sociaux ont été largement étudiés par les sociologues, les anthropologues et les psychologues. Pour eux, un rôle social est un comportement qu'une communauté attend d'une personne qui occupe une position dans cette communauté. Une étude ethnologique canonique des rôles dans les forums en ligne a été faite dans Golder (2003). Les rôles en ligne ont également été étudiés par des informaticiens, qui ont mis davantage l'accent sur la détection des rôles. En informatique, un rôle est généralement considéré comme un ensemble de fonctions centrées sur l'utilisateur ou la position que l'individu détient dans le graphe social.

Nous pensons que l'un des aspects les plus intéressants du rôle dans la sociologie est que, une fois que nous savons le rôle d'un individu, nous pouvons prédire, dans une certaine mesure, la façon dont l'individu va se comporter dans une certaine situation. Les rôles sont des catégories descriptives et prédictives du comportement. Nous savons comment traiter les deux problèmes séparément : une catégorisation descriptive des utilisateurs est une tâche d'apprentissage non supervisé (clustering), tandis qu'apprendre à prédire les comportements des utilisateurs à partir d'un historique des comportements passés est souvent une tâche d'apprentissage supervisé. Cependant, intégrer à la fois ces deux tâches n'est pas évident.

L'objectif de cette thèse est précisément d'intégrer à la fois une vision *descriptive*

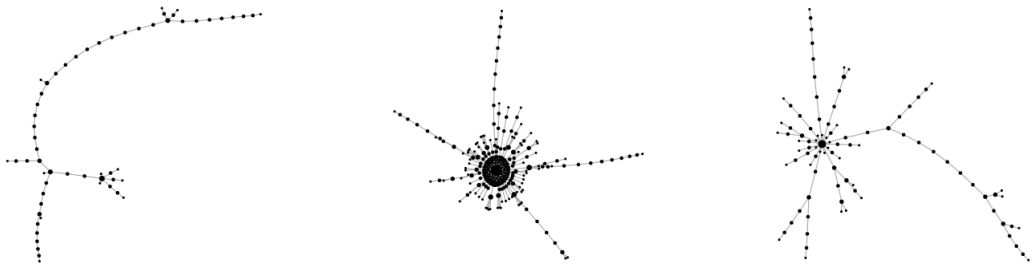


Figure 1: Conversation trees with different structures. Different user roles are assumed to contribute differently to the growth of the trees.

et *prédictive* des rôles en ligne.

En imaginant les rôles comme des comportements typiques, nous allons explorer quelques façons de trouver des groupes d'utilisateurs ayant des comportements similaires.

1.2 De nouvelles approches pour la détection des rôles

Les méthodes actuelles de détection de rôle sont basées sur l'analyse des interactions à partir de différents points de vue (blockmodels, caractéristiques ou triades). Dans cette thèse, nous descendons au niveau conversationnel, représenté par les arbres de messages (posts), pour trouver des rôles basés sur les différentes manières dont les gens se parlent lorsqu'on prend en compte la structure.

Dans le chapitre 3, nous présentons notre première méthode pour la détection des rôles basée sur les structures conversationnelles. Nous appliquons différentes notions de voisinage pour les messages dans les graphes d'arbres (*radius-based*, *order-based*, and *time-based*) et nous montrons que nos voisinages *order-based*, de manière similaire aux triades dans d'autres types de graphes, sont en mesure de capturer la plupart des structures conversationnelles pertinentes sans avoir besoin de motifs complexes. Nous utilisons ces voisinages pour détecter des groupes d'utilisateurs qui ont une tendance à participer au même type de conversation.

Dans le chapitre 4, nous présentons notre deuxième méthode basée sur des modèles stochastiques de croissance pour les fils de conversation. En utilisant des modèles génératifs, nous proposons une méthode pour trouver des groupes d'utilisateurs qui ont tendance à répondre au même type de messages. Nous montrons que, alors que nous sommes en mesure de trouver des groupes d'utilisateurs en fonction de leurs comportements passés, il n'y a aucune preuve que ces comportements soient prédictifs des comportements futurs. Cette découverte remet en question notre conception de la notion de *rôle* comme un comportement cohérent d'un utilisateur, au moins en ce qui concerne le type de comportement conversationnel que nous considérons ici.

Notez que, dans le chapitre 3, nous regroupons les utilisateurs en fonction de vecteurs de caractéristiques (*feature vectors*) qui sont directement construits à partir des comportements observés et, dans le chapitre 4, nous supposons une fonction de comportement sous-jacente avec les paramètres latents et nous regroupons les utilisateurs en fonction de leurs paramètres estimés. Dans le chapitre 5, nous présentons une troisième méthode qui intègre le type de données utilisées dans

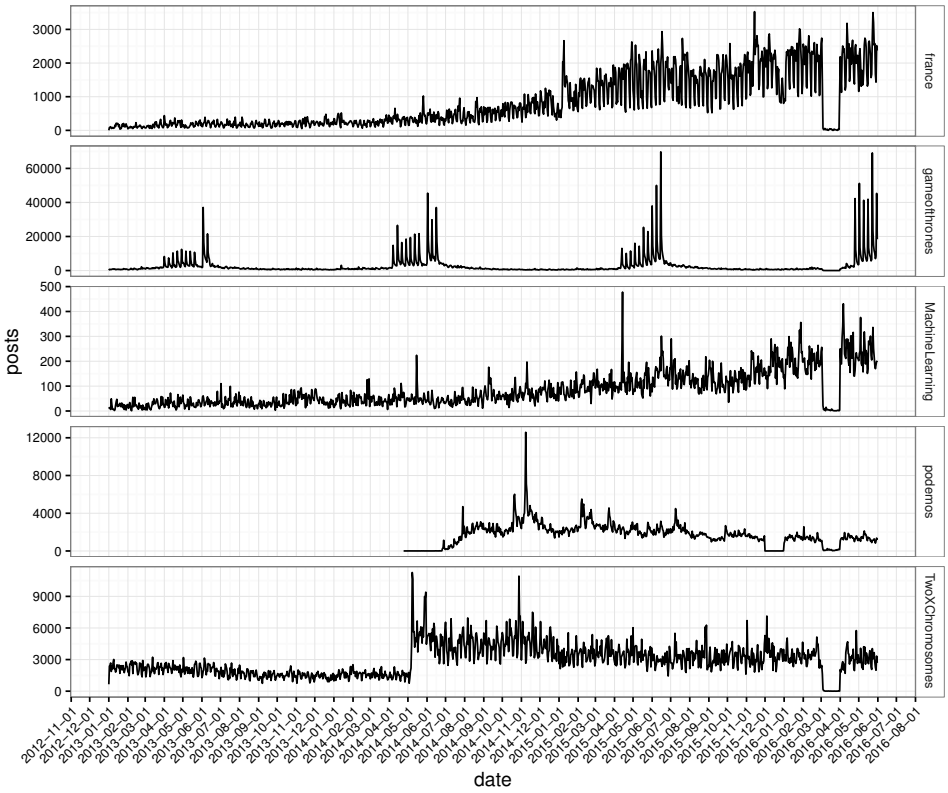


Figure 2: Posts by day in the analyzed forums

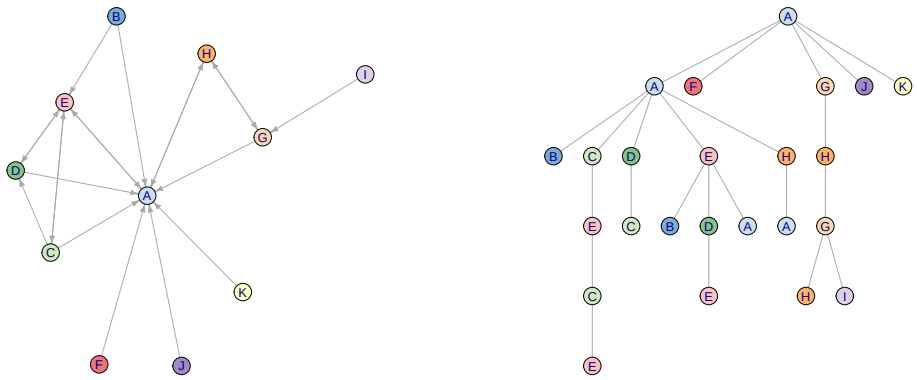


Figure 3: A thread represented as a graph of interactions and as a tree graph.

les deux chapitres précédents (*feature-based* et *behavioral* ou *functional-based*) et qui peut trouver des groupes en utilisant moins d'exemples. Le modèle exploite l'idée selon laquelle les utilisateurs ayant des caractéristiques similaires ont des comportements similaires. Non seulement la méthode intègre des entrées de nature différente, en ajoutant plus de cohérence aux clusters, mais il aborde le problème du clustering des utilisateurs lorsque la plupart des utilisateurs ont seulement participé à quelques reprises. Nous utilisons des données synthétiques pour afficher les propriétés de ce modèle tout en laissant son application aux forums réels pour des recherches futures.

Nous concluons ce manuscrit au chapitre 6 avec un résumé des résultats et une discussion sur les perspectives ouvertes par cette thèse.

2 Détection de rôles basée sur des structures de conversation

Dans ce chapitre, nous étudions la pertinence des motifs de graphes pour détecter différents profils de discutants. Nous représentons un fil comme un arbre de messages où chaque message a un auteur. Le voisinage d'un message est un sous-graphe induit qui contient la partie locale de l'arbre qui se trouve autour. Ainsi, le voisinage capture la structure de la conversation locale où le message est intégré. Comme nous le verrons dans ce chapitre, il n'y a pas de définition "naturelle" du voisinage dans le contexte des arbres de conversation. Nous proposerons trois définitions alternatives avant d'analyser leurs forces et leurs faiblesses.

Par conséquent, nous allons regrouper les utilisateurs en fonction du type de voisinage où ils ont tendance à apparaître. Nous allons présenter et analyser trois nouvelles définitions de voisinage : *radius-based, order-based* et *time-based*. Notre stratégie pour détecter les différents types de conversateurs est la suivante : d'abord, nous extrayons le voisinage du message de chaque utilisateur ; deuxièmement, nous regroupons les utilisateurs en fonction de leur tendance à apparaître dans chaque type de voisinage.

Nous comparons les trois définitions différentes du voisinage pour capturer la structure dans laquelle un message est intégré. Le choix du voisinage approprié n'est pas trivial à cause d'un compromis entre l'expressivité (c'est à dire, la variété des structures qu'une définition de voisinage capture) et la sparsity du dictionnaire (c'est à dire, le nombre de classes de voisinage, ou des motifs, qui ont des fréquences très basses).

Pour différencier un peu mieux les messages, par exemple entre une réponse commune et une réponse à un message situé à la racine, nous avons utilisé des couleurs correspondant aux messages écrits par l'utilisateur *ego* et au message racine. Nous appliquons également une stratégie d'élagage pour enlever les parties redondantes des conversations qui n'ajoutent aucune information supplémentaire à priori.

En observant la taille réduite de son dictionnaire et le type de conversations qu'il est capable de détecter, nous considérons que le *order-based* est le plus prometteur parmi tous les voisinages proposés. Une éventuelle amélioration consisterait à fusionner manuellement –probablement sous certains critères – les motifs dont la différence ne paraît pas significative d'un point de vue sociologique.

Dans l'ensemble, nous montrons que la structure de la conversation fournit de nouvelles informations qui ne sont pas dévoilées par d'autres méthodes basées sur des caracteristiques. Nous pensons que la structure de cluster que nous observons clairement dans certains cas (à savoir : des utilisateurs ont une préférence très

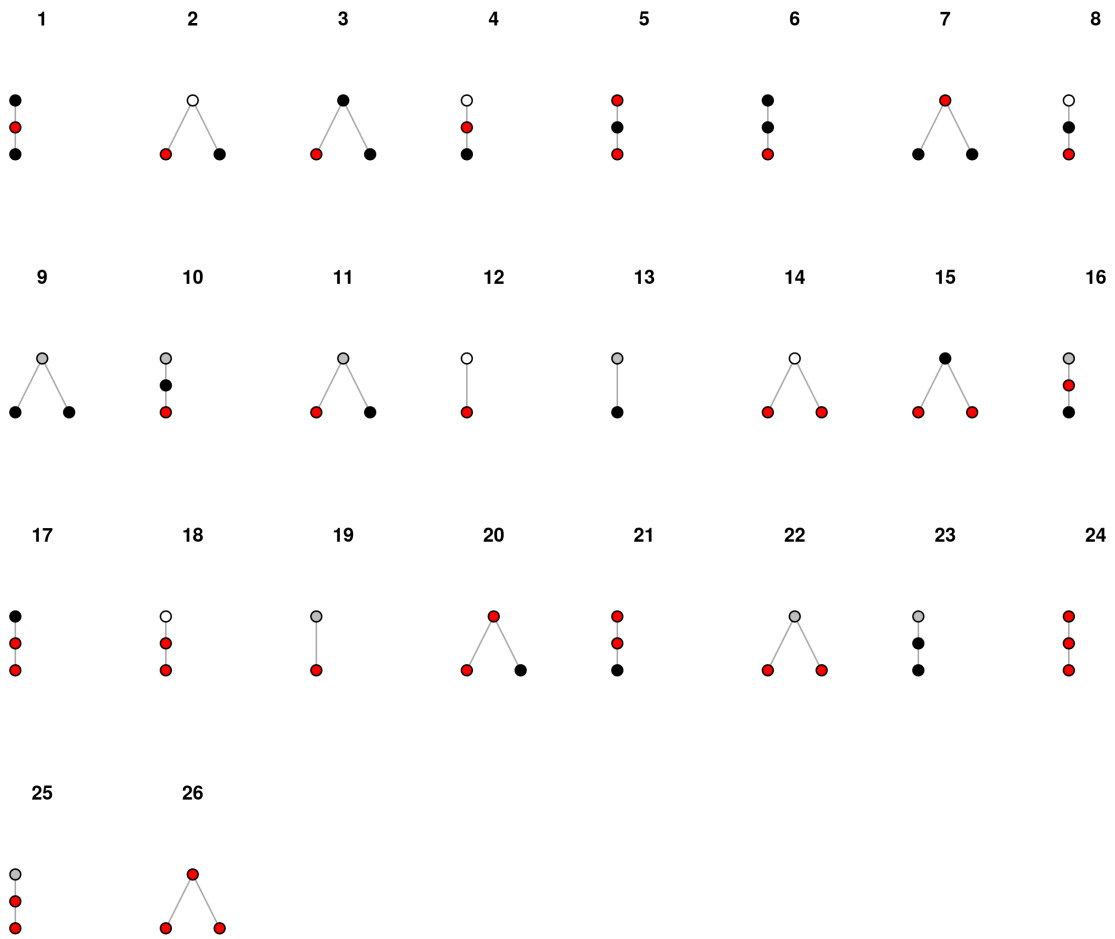


Figure 4: Aperçu des motifs conversationnels les plus fréquents, selon notre définition de voisinage order-based.

marquée vers un certain type de structures) soutient cette affirmation.

3 Détection de rôles basée sur des modèles de croissance de graphes

Dans le chapitre précédent, nous proposons de détecter les rôles sur la base de la structure des conversations auxquelles les utilisateurs participent. Cette méthode peut être un bon complément à une autre méthode basée sur les caractéristiques car il prend en compte une nouvelle dimension du comportement des utilisateurs : les conversations. Néanmoins, cette approche est purement descriptive. Il y a encore un chaînon manquant qui relie le comportement observé d'un utilisateur avec une certaine *fonction comportementale* qui modélise la raison pour laquelle l'utilisateur a choisi telle ou telle action. On conceptualise une fonction comportementale comme une distribution de probabilité sur l'espace de tous les comportements possibles dans un contexte donné. Nous supposons qu'il existe un répertoire fini de fonctions comportementales et que tous les comportements observés d'un utilisateur sont tirés à partir d'une de ces fonctions. Nous établissons que les utilisateurs qui partagent

la même fonction comportementale tiennent le même rôle. Ceci étant dit, notre définition de rôle dans ce chapitre est la suivante :

> Deux utilisateurs jouent un même rôle à partir du moment où ils ont tendance à partager les mêmes (paramètres d'une) fonction comportementale.

Dans ce chapitre, nous établissons trois objectifs principaux : (a) proposer une fonction comportementale pour les fils de discussion, (b) trouver des groupes d'utilisateurs avec la même fonction comportementale (les mêmes paramètres), et (c) vérifier si ces fonctions comportementales ont un pouvoir prédictif — si elles permettent de prédire le comportement d'un utilisateur dans un nouveau contexte.

Nous utilisons des modèles de graphes aléatoires comme base pour nos fonctions comportementales. En particulier, nous allons nous intéresser aux modèles de croissance. Les modèles de croissance sont des générateurs aléatoires de graphes qui tentent d'imiter le mécanisme de croissance d'un réseau à travers des processus stochastiques régis par un ensemble de paramètres. Formellement, un modèle de croissance définit une distribution de probabilité qui quantifie la probabilité pour un sommet existant i d'être choisi comme parent pour un nouveau sommet x_t dans le graphe :

$$p(x_t \sim i | G_{t-1}; \boldsymbol{\theta}) \tag{1}$$

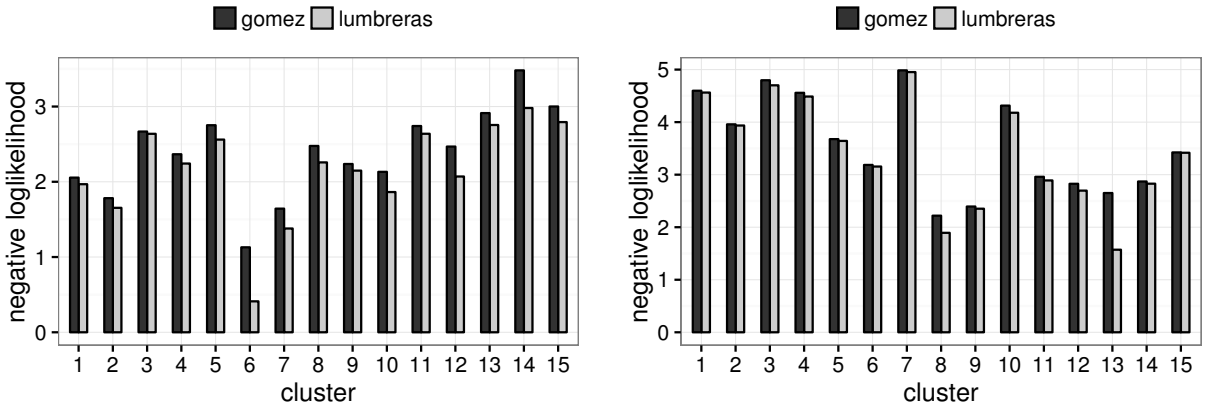
où G_{t-1} est l'état du graphe avant l'arrivée de x_t et $\boldsymbol{\theta}$ sont les paramètres du modèle — la spécification de cette distribution de probabilité dépend de quelle hypothèse raisonnable nous pensons pouvoir suivre pour le processus de croissance. Ces modèles peuvent être considérés comme des fonctions comportementales car ils modélisent la façon dont les utilisateurs choisissent le message auquel ils vont répondre. Le répertoire de comportements possibles est alors un ensemble de paramètres $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$, et la probabilité ci-dessus dépendra donc du $\boldsymbol{\theta}$ associée à l'auteur du message x_i — correspondant au rôle de l'auteur.

Afin de résoudre notre problème, nous proposons un modèle de croissance du fil de discussion qui permet aux messages écrits par différents utilisateurs d'être associés à des paramètres de croissance différents $\boldsymbol{\theta}$. L'idée est très simple et consiste à estimer, par Expectation-Maximization, des groupes d'utilisateurs ayant leurs propres paramètres.

Nous montrons que, en effet, nous pouvons trouver différents groupes d'utilisateurs ayant des fonctions différentes de comportement. Cela signifie que notre modèle peut être utilisé, par exemple, pour mieux comprendre la dynamique d'une communauté en inférant différents groupes d'utilisateurs qui ont contribué à cette dynamique de différentes manières.

En ce qui concerne le pouvoir prédictif, nous utilisons les différents paramètres du modèle appris afin de tester s'il est possible d'inférer des comportements sur des données non utilisées pour l'estimation. Néanmoins, cette amélioration de la probabilité ne suffit pas pour faire de meilleures prévisions. En effet, en termes de prédictions pratiques associées à la question *quel message sera le prochain à être répondu ?*, notre modèle basé sur les rôles ne fait pas de meilleures prédictions — pour la plupart des rôles — qu'un modèle qui ne se base pas sur les rôles.

Sur cette base, notre conclusion est que notre concept proposé de rôle de comportement a un certain pouvoir descriptif mais que son pouvoir prédictif reste marginale. Il se pourrait que la cohérence des comportements soit en effet faible — mais pas totalement aléatoire — et que, en termes de signal, il y a trop de *bruit* pour obtenir des résultats fiables. Il se pourrait aussi que les modèles de croissance



(a) `podemos` (b) `gameofthrones`

Figure 5: Moyenne de la Log-vraisemblance négative par cluster dans l'ensemble de test (valeur à minimiser)

des arbres présentés dans ce chapitre ne soient pas en mesure de capturer la petite partie de ce signal de comportement.

4 Détection de rôles basée sur des caractéristiques et des fonctions comportementales latentes

Un problème récurrent que nous avons observé dans les chapitres précédents est la rareté des données pour un même utilisateur. Avant de pouvoir affirmer qu'un utilisateur appartient à un groupe donné, il est nécessaire d'avoir accès à un volume suffisant de données de l'utilisateur, de sorte que nous pouvons être sûrs que ses comportements observés sont vraiment liés au rôle qu'il joue et non pas seulement à un comportement occasionnel sans aucune signification pertinente. Pour surmonter ce problème, nous proposons, dans ce chapitre, un modèle dual-view qui intègre les deux approches précédentes.

Les modèles dual-view regroupent les utilisateurs en fonction des caractéristiques observées et des fonctions comportementales latentes afin d'inférer des groupes d'utilisateurs qui sont plus robustes et significatifs. L'objectif est d'en déduire la fonction du comportement de chaque utilisateur mais également un regroupement des utilisateurs qui tient à la fois compte de leurs caractéristiques et de leurs fonctions comportementales. Dans notre modèle, les utilisateurs du même groupe sont supposés d'avoir des caractéristiques similaires *et* fonctions comportementales, et donc l'inférence des clusters dépend de l'inférence des fonctions comportementales, et vice versa. Le modèle priorise les partitions d'utilisateurs où chaque groupe possède à la fois des caractéristiques similaires *et* des fonctions comportementales similaires.

Nous testons le modèle sur des données synthétiques afin d'évaluer et de comprendre ses propriétés. Les caractéristiques, comme les coefficients de comportement dans nos données, sont tirées de distributions gaussiennes.

Le modèle est un modèle de mélanges probabiliste qui regroupe les utilisateurs en fonction des caractéristiques et des fonctions comportementales latentes (Figure 6).

Chaque composante du modèle de mélange représente une densité de probabilité

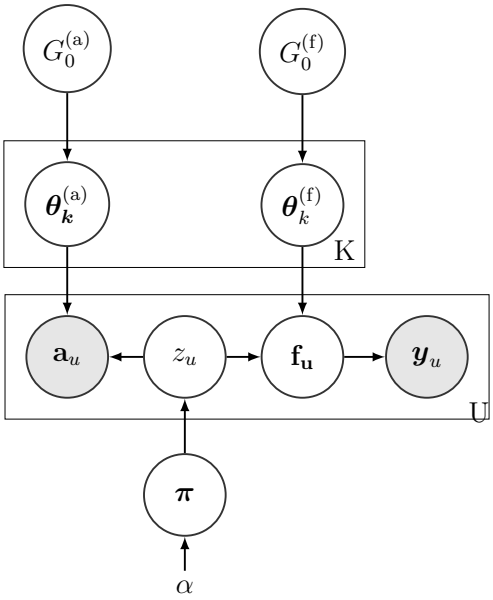


Figure 6: Modèle graphique du processus génératif pour U utilisateurs et K clusters. Les deux vues sont connectées par le biais des assignations latentes $\mathbf{z}$. Un vecteur de caractéristiques $\mathbf{a}_u$ et un comportement $\mathbf{f}_u$ sont générés pour l'utilisateur u à partir de son cluster, indiqué par z_u.

sur deux vues : une vue pourles attributs observés (ex. : degré dans le graphe, centralité) et une vue pour les fonctions comportementales latentes qui sont indirectement observés par des actions ou des comportements des utilisateurs (ex. : générer de longues conversations). La distribution postérieure représente un regroupement qui réalise un compromis entre les deux points de vue. L'inférence des paramètres de chaque vue dépend de l'autre vue à travers ce regroupement commun qui peut être considéré comme un proxy transmettant les informations entre les deux points de vue. Une propriété intéressante du modèle est que l'inférence sur les fonctions comportementales latentes peut être utilisé pour faire des prédictions des utilisateurs comportements futurs. Nous présentons deux versions du modèle : une version paramétrique, où le nombre de clusters est traité comme un paramètre fixe, et une version non paramétrique, basée sur un processus de Dirichlet, où le nombre de groupes est également déduit automatiquement.

Nous adaptons le modèle à un cas hypothétique de forums en ligne où les comportements correspondent à la capacité des utilisateurs de générer de longues discussions. Nous regroupons les utilisateurs et déduisons leurs fonctions comportementales dans trois ensembles de données pour comprendre les propriétés du modèle. Nous déduisons les probabilités à postériori d'intérêt par un échantillonnage de Gibbs sur l'ensemble des variables, sauf pour deux d'entre elles qui sont inférées par Adaptive Rejection Sampling. Les expériences confirment que le modèle dual-view proposé est capable d'apprendre avec moins d'observations que sa contrepartie single-view, et ce en raison du fait que le double-view a accès à davantage d'informations. En outre, les inférences avec le modèle dual-view sur la base d'un processus de Dirichlet sont aussi bonnes que les inférences avec le modèle paramétrique, même si celui-ci connaît le vrai nombre de groupes.

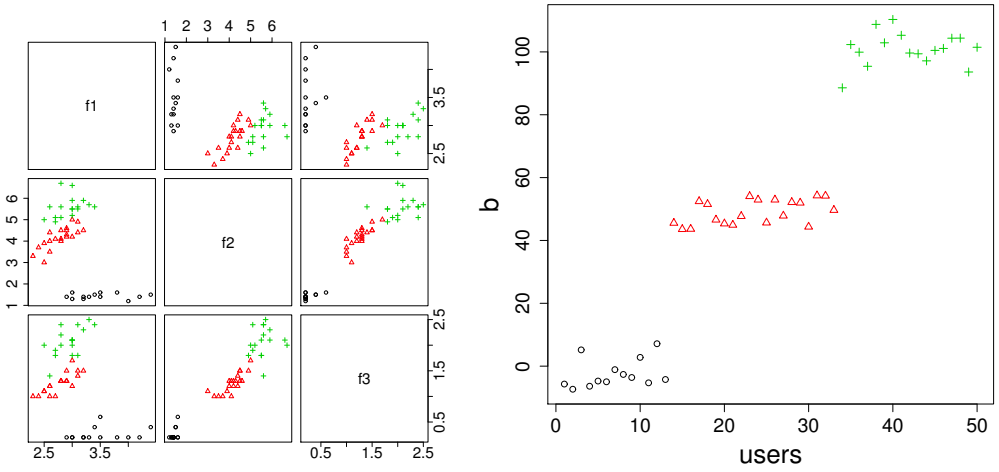


Figure 7: Variables décrivant les utilisateurs (à gauche) et les coefficients de la fonction de comportement (à droite), données artificielles utilisées pour tester le modèle

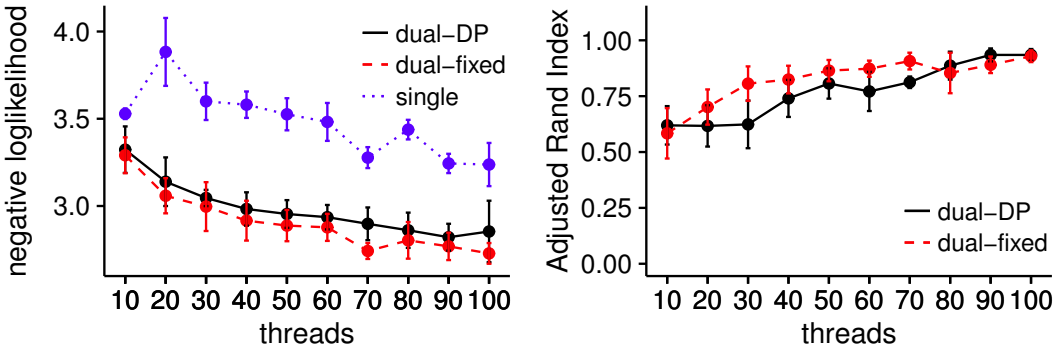


Figure 8: Résultats expérimnetaux sur l'analyse du nombre de fils de discussion nécessaires, par utilisateur, pour obtenir de bonnes performances (moyenne et écart-type calculés sur 5 exécutions)

References

Duggan, M. (2015). Mobile Messaging and Social Media 2015. Technical Report August, Pew Research Center.

Golder, S. A. (2003). *A Typology of Social Roles in Usenet*. Ph. D. thesis, Harvard University.

Gómez, V., H. J. Kappen, N. Litvak, and A. Kaltenbrunner (2012, apr). A likelihood-based framework for the analysis of discussion threads. *World Wide Web 16*(5-6), 645–675.

Kollock, P. (1998). *Communities in Cyberspace*. London: Routledge.

Kumar, R., M. Mahdian, and M. McGlohon (2010). Dynamics of Conversations. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 553–562.

McFarland, D. A., K. Lewis, and A. Goldberg (2016). Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *American Sociologist 47*(1), 12–35.

Tinati, R., S. Halford, L. Carr, and C. Pope (2014). Big Data: Methodological Challenges and Approaches for Sociological Analysis. *Sociology 48*(4), 663–681.