Automatic role detection in online forums

Soutenance de thèse

Alberto LUMBRERAS CARRASCO

Directeurs:

Bertrand JOUVE
Julien VELCIN
Marie GUÉGAN (Technicolor)

Novembre 7, 2016

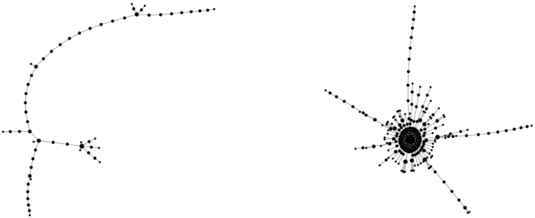
Roles

Parent, son, friend, doctor, engineer... **behaviors** *often attached to social positions.*

Useful because they are a mean to understand individual (and even collective) behavior.

Motivation

Roles and structural dynamics



- How different roles contribute to the **structure of conversations**?
- Can we use roles to model (and predict) user behaviors?

Industrial context

technicolor

- Media and entertainment sector, film industry.
- Growing attention to **end-users**: user profiling, recommender systems,...

Why **forums**?

- Reaction to movies, discussion about particular scenes,..
- Previous in-house work with Internet Movie Database (IMDb).

Outline

Outline

Roles in sociology

- **No universal definition**. Many approaches in sociology, influenced by the different schools (structuralism, symbolic interactionism, functionalism,...).

An attempt to look for the **common denominator**[1]:

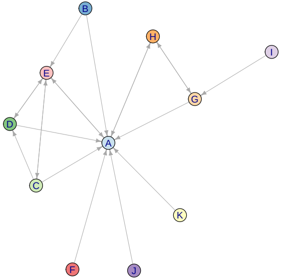
> *"In current social science the term role has come to mean a* **behavioral repertoire characteristic of a person or a position***; a set of standards, descriptions, norms, or concepts held for the behaviors of a person or social position; or (less often) a position itself."*

[1] Bruce J Biddle. *Role Theory: Expectations, Identities, and Behaviors*. New York: Academic Press, 1979.

Dual representation

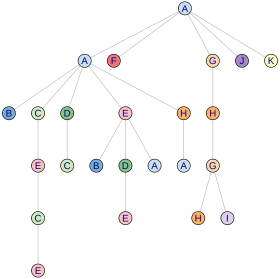
Social Network representation

- Focus on **social** structure
- Positions, centrality, cliques...



Tree of posts representation

- Focus on **conversation** structure.



Blockmodeling

Roles as positions in social structure

- Finds a relational structure in an adjacency matrix.
- **Positions** in the structure are often related to **roles**.

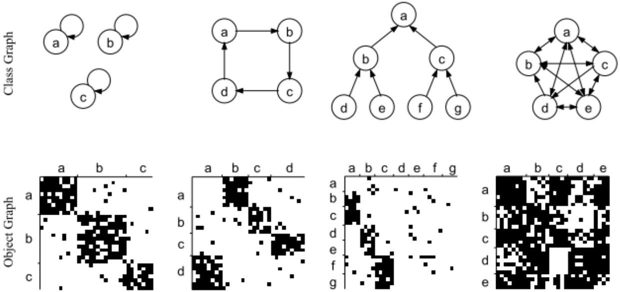


Figure : Stochastic Blockmodeling over different social structures.[2]

[2] Charles Kemp, Thomas L Griffiths, and Joshua B Tenenbaum. *Discovering latent classes in relational data*. Tech. rep. Massachusetts Institute of Technology, 2004.

Feature-based

Roles as sets of features

- Centrality measures, #posts[3], #threads started, #votes/post[4], # posts with reply, mean posts/thread[5], clustering coefficient in social neighborhood[6], ...
- Clustering over selected features.

---

[3] Mathilde Forestier et al. "Extracting celebrities from online discussions". In: *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*. 2012, pp. 322–326.

[4] Matthew Rowe et al. "Community analysis through semantic rules and role composition derivation". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 18.1 (2013), pp. 31–47.

[5] Jeffrey Chan, Conor Hayes, and Elizabeth Daly. "Decomposing discussion forums using common user roles". In: *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. 2010.

[6] Cody Buntain and Jennifer Golbeck. "Identifying Social Roles in Reddit Using Network Structure". In: *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. 2014, pp. 615–620.

Triad-based

Role as distributions over triads

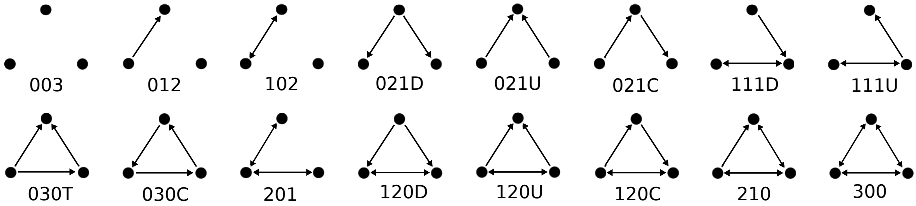


Figure : List of all possible triads.

- Count number of times a user appears in each triad.
- $\mathbf{f}_u = (\%t_1, ..., \%t_{16})$
- Clustering over vector of counts $\mathbf{f}_1, ..., \mathbf{f}_U$.

Pros and cons

Blockmodeling

- Pros: Sociologically grounded.
- Cons: In forums, positions $\sim$ behavior less clear than in stable social structures.

Feature-based

- Pros: Easy, fast, transparent.
- Cons: Arbitrary selection of features.

Triads

- Pros: Common tool in biology, SNA,...
- Cons: Cyclic graphs not adapted to trees.

And none of them have predictive power.

Roles based on conversation structures

Role $\rightarrow$ behavior $\rightarrow$ conversation

Role detection based on a basic form of behavior in *discussion* forums: conversations.

Conversational-based roles.

- *motif-based*: in what structural kind of conversation does the user participate?
- *function-based*: what is the behavioral function of a user?

Combining features and functions:

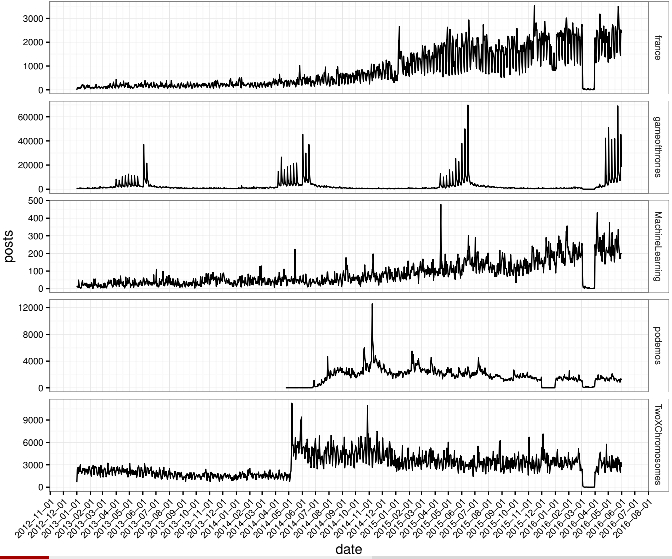
- *feature + function*: how can we detect roles based on both feature/motifs and functional descriptions of behaviors?

The data

Reddit. A forum of forums

- 2013-2016.

MachineLearning
Podemos
France
TwoXChromosomes
GameofThrones



Unbalanced user activity

Unbalanced user participations. Roles might alleviate this problem by extrapolation.

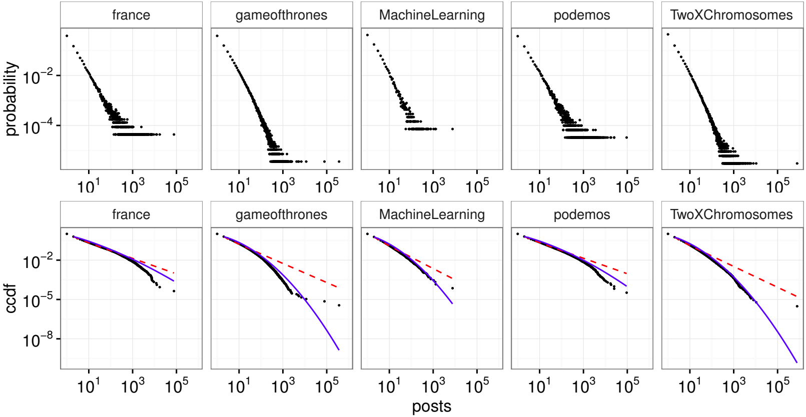


Figure : Number of posts (PDF and CCDF). MLE fits of Power Law (dashed) and Log-normal (solid) distributions.

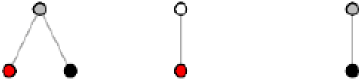
Outline

Idea

You are the way you structurally talk

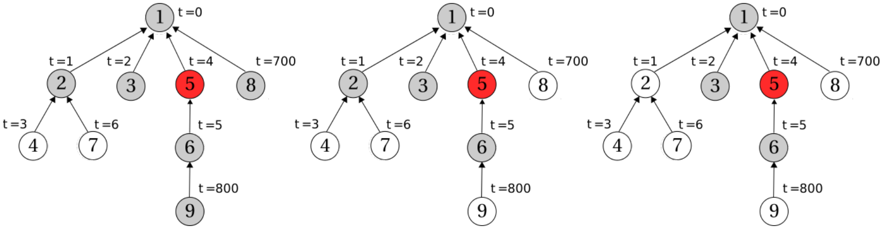
Definition

Two users have the same role if they tend to participate in the same positions of the same type of neighborhoods (motif)



Neighborhoods

Radius/Time/Order-based neighborhoods



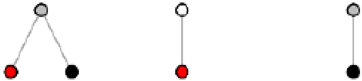
(a) Radius-based $r = 2$ (b) Time-(radius)-based $r = 2$ (c) Order-based $o = 4$

- **Radius-based**: every post at *distance* $d \leq r$ from ego.
- **Time-based**: every post at *distance* $d \leq r$ from ego and before speed *changepoints*.
- **Order-based**: o posts *closest* in time to ego (including ego).

Coloring and pruning

Colors to identify the type of post:

- root: white
- ego post: red
- ego + root: grey
- other: black



Pruning to avoid large neighborhoods:

- Allow only two replies with the same color.

Methodology

1. Neighbourhood extraction
   - Radius-based extraction
   - Order-based extraction
   - Time-based extraction
2. Clustering
   - Hierarchical clustering (cut at height $h = 10$)
   - but other methods are also possible

Our aim

Compare radius-based, order-based, time-based.

- How many motifs?
- What conversations do they represent?
- What types of users do they discover?

Motif selection

- Compute probability of each user to appear in each motif/neighborhood: $\mathbf{f}_u = (p_{u1}, ..., p_{uN})$ (*see figure*)
- Compute median per motif (red dots).
- Remove those with median 0 unless a 10% of outliers (Tukey's test).

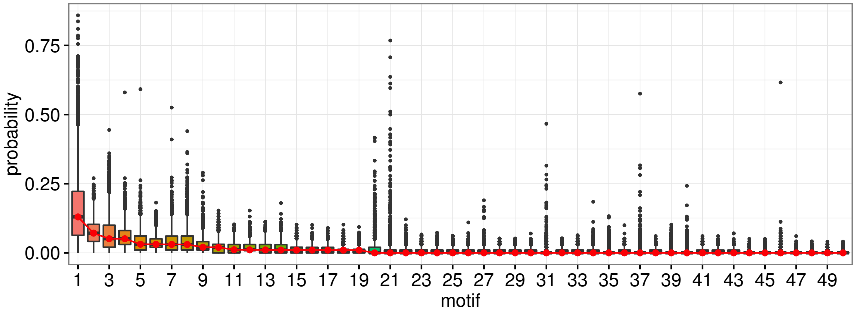


Figure : Probability of motif by user (radius-based $r = 2$)

Size of dictionary

neighborhood	dictionary	selected features
radius $r = 2$	1269	27 (19 + 8)
time $r = 2$	746	23 (19 + 4)
order $o = 3$	26	17 (7 + 10)

Order-based $o = 3$

- **26 motifs.** 7 motifs with median $> 0 + 10$ with more than 10% outliers $=$ **17 features** (selected motifs) (6 less than time-based)

Order-based makes a better use of the motifs space.

Expressiveness

Order-based $o = 3$

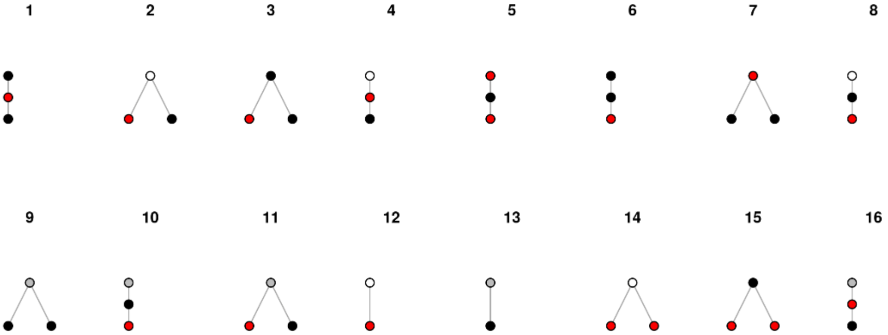


Figure : Dictionary of the first motifs sorted by median probability

- Able to capture: Cascades, multiple replies, terminations...

Clusters

Order-based $o = 3$

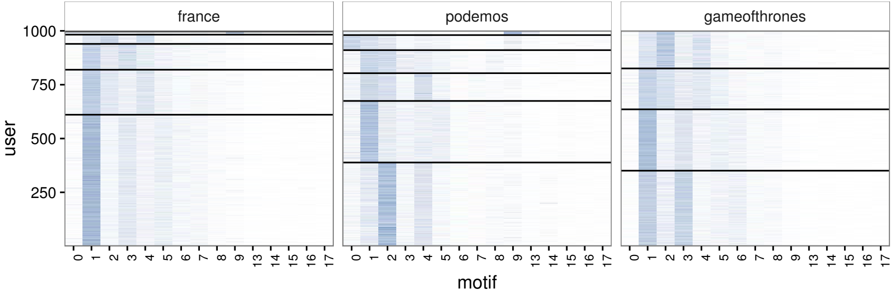


Figure : Clusters with order-based neighborhoods (r=3)

Roles

Order-based $r = 3$

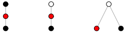
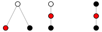
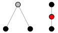
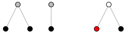
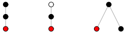
Role	main motifs	ID motifs	forums
Successful repliers		1,3,*	fr, got
Successful repliers		1,4,2	fr, pod
Successful repliers		1,2,*	pod, got
root repliers		2,4,1	fr, pod, got
initiators	 ?	9,1,0	fr
initiators		9,13,2	pod
terminators		6,8,3	fr
others	? 	0,1,2	pod

Table : Summary of clusters with order-based neighborhoods (o=3). Clusters with similar first and second motif, but different third motif, have been collapsed into a same group. The question mark corresponds to the *others* category.

Discussion

- Size of dictionary:
  - Order-base makes a better use of the feature space.
- Expressiveness:
  - All methods are able to capture cascades, stars...
  - But big dictionaries are too sensitive (very similar conversations represented by different motifs).
- Clustering and roles:
  - Although all detect meaningful clusters, radius-based and time-based do not use most of the features.

Order-based is the most promising neighborhood.

Possible improvement: manually merge order-based motifs that represent similar conversations.

Outline

Generative models

- Graph generative processes that account for some relevant properties of real graphs (and the simpler, the better!).

Preferential Attachment: $p(x \sim i) \propto d_i^{\alpha}$ *"Rich get richer"*

Figure : $\alpha = 0.1$

Figure : $\alpha = 1$

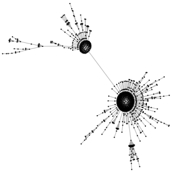
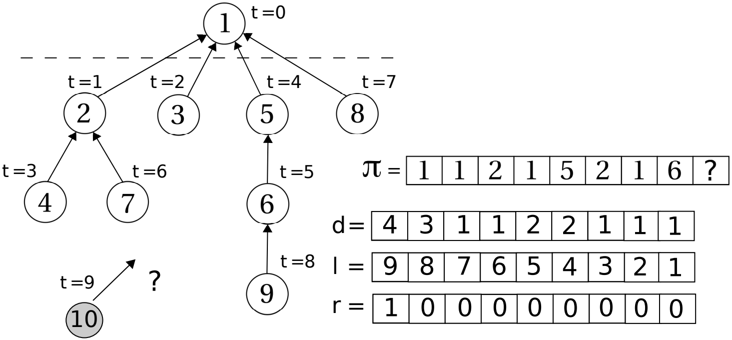


Figure : $\alpha = 1.5$

Thread growth models

Modeling the evolution of trees



d_i: popularity (degree); r_i: root or not; l_i: recency

- Modeling the choices.
    - Barabasi: $p(\pi_t \sim i) \propto d_i^{\alpha}$
    - Kumar 2012: $p(\pi_t \sim i) \propto \alpha d_i + \tau^{l_i}$
    - **Gomez 2012**: $p(\pi_t \sim i) \propto \alpha d_i + \beta r_i + \tau^{l_i}$

Role detection based on thread growth

Idea: You are the way you choose whom to reply

- Current models estimate the same parameters for all users.
- Idea:
  - *Gómez 2012*[7] as base model: $p(\pi_t \sim i) \propto \alpha d_i + \beta r_i + \tau^{l_i}$
  - Estimate different parameters for different users, allowing different behaviors.

$$p(\pi_t \sim i) \propto \alpha_{z_u} d_i + \beta_{z_u} r_i + {\tau_{z_u}}^{l_i}$$

We will say that *two users have the same role if they have the same parameters of their behavioral function.*

[7] Vicenç Gómez et al. "A likelihood-based framework for the analysis of discussion threads". In: *World Wide Web* 16.5-6 (2012), pp. 645–675. arXiv: 1203.0652.

Role detection based on thread growth

Formalization

Log-likelihood given cluster assignments $\mathbf{Z}$ and parameters $\boldsymbol{\theta}$:

$$\ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta}) = \sum_{u=1}^{U} \sum_{n \in N_u} \ln \left(\alpha_{z_u} d_n + \beta_{z_u} r_n + \tau_{z_u}^{l_n} \right) - \ln Z_n \tag{1}$$

The model naturally fits an E-M algorithm.

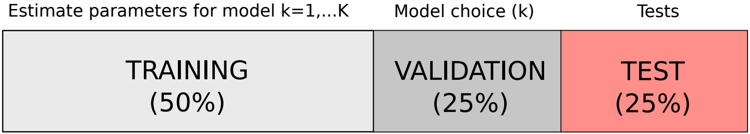
$$\overbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \underbrace{\left(\ln p(\mathbf{Z}|\boldsymbol{\pi}) + \ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta})\right)}_{\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}}^{\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})]} \tag{2}$$

- *Expectation*: Re-compute cluster assignments $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$.
- *Maximization*: Maximize Eq. 2 w.r.t cluster parameters $\boldsymbol{\theta}, \boldsymbol{\pi}$ (Nelder-Mead optimisation).

Experiments

Setting

- **Forums**: Game of Thrones (major results similar for all forums).
- For each of the top 1000 most active users:



- **Training**: estimation of parameters.
- **Validation**: choice of number of clusters with BIC.
- **Test**: predictions.

Experiments

Estimate parameters

- Our role-based growth model allows more flexibility to detect **outlier** behaviors.

cluster	α	β	τ	π	users
1	0.1	0.66	0.96	0.08	89
...	...	...	....	....	...
8	0.01	**81.89**	0.98	0.03	26
9	0.03	2.84	0.8	0.08	77
10	0	4.12	0.99	0.04	39
11	0.4	12.16	0.95	0.02	19
12	0.07	9.05	0.85	0.05	54
13	0	**0.1**	**0.43**	0	8
14	0.02	0.93	0.76	0.13	128
15	0.06	5.13	0.96	0.12	120
Gomez	**0.06**	**2.71**	**0.93**	-	

Experiments

Generated threads

- Role-based model generate similar threads to Gomez

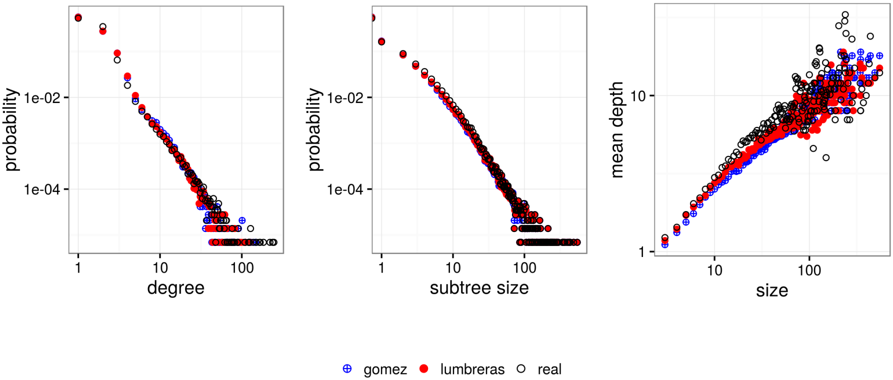
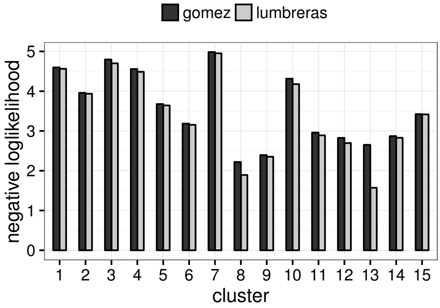


Figure : Properties of synthetic trees and real trees

Experiments

Link prediction. Predicting the choices of parent in the test set.

Role-based model outperforms especially in outlier clusters:



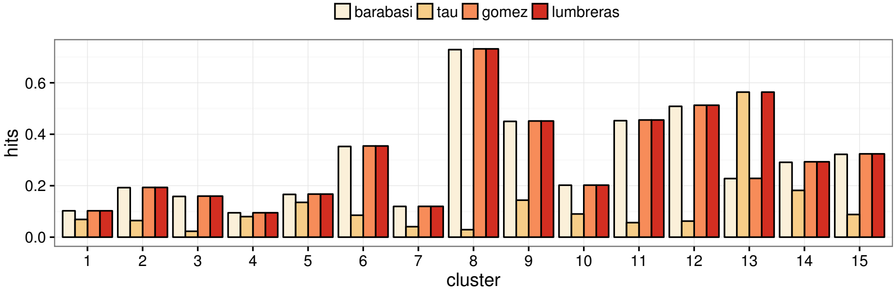
cluster	α	β	τ	users
8	0.01	**81.89**	0.98	26
13	0	**0.1**	**0.43**	8
Gomez	**0.06**	**2.71**	**0.93**	-

Experiments

Link prediction. Predicting the choices of parent in the test set.

Compared models:

- *Barabasi*: always replies to the post with more replies.
- *Tau*: always replies to the most recent post.
- *Gomez*: $p(\pi_t \sim i) \propto \alpha d_i + \beta r_i + \tau^{l_i}$
- *Lumbreras*: Gómez with one set of parameters per cluster.



Lumbreras $\succ$ Gomez in **cluster 13**.

Discussion

- A growth model $p(\pi_t \sim i) \propto \alpha d_i + \beta r_i + \tau^{l_i}$
- with different α, β, τ for every cluster (role).
- in order to detect the latent roles and their behavioral parameters.

Clustering as a way to understand and categorize users according to their behaviors.

- Detection of groups of users that behave differently.

Predictions as a validation test for the **existence of roles**.

- Role-based model improves likelihood of new observations, specially for outlier clusters.
- **But** likelihood is not higher enough to make a difference in predictions (except for some **outliers** with extreme behaviors).

Either the role signal is weak or we need a better growth model.

Outline

Why a dual-view model?

We want to integrate two types of features:

- observed features (e.g.: motifs frequence)
- latent behavioral functions (α, β, τ of previous model)

Besides, lots of users with low activity :

- For users with a few posts, not enough information to be confident about their parameters.

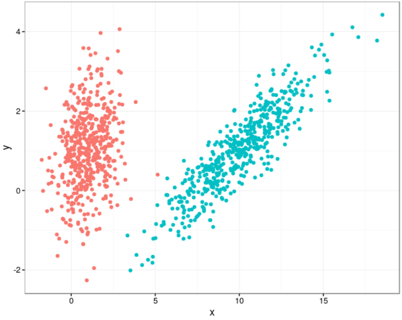
Key idea:

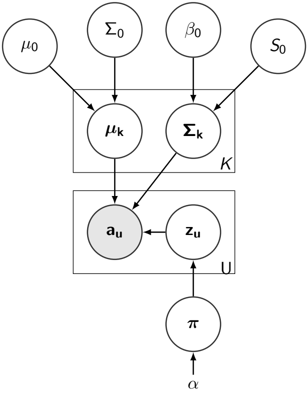
- If users with similar features have similar behavioural parameters, then we can *cheat* using this information to help inference of behavioural parameters.

Mixture models

Gaussian Mixture Model

$$\pi \sim \text{Dirichlet}(\boldsymbol{\alpha})$$
$$z_i \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{\Sigma}_k \sim \mathcal{W}(\beta_0, \mathbf{S}_0)$$
$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$
$$a_u | z_i, \boldsymbol{\mu}_{z_u}, \boldsymbol{\Sigma}_{z_u} \sim \mathcal{N}(\boldsymbol{\mu}_{z_u}, \boldsymbol{\Sigma}_{z_u})$$



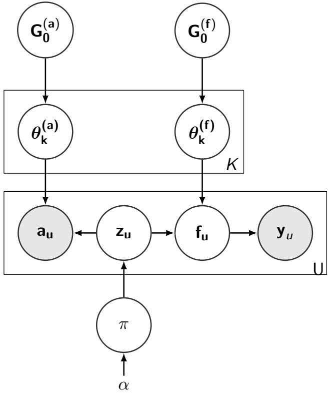


Dual-view mixture model

Features view + behaviors view

Two views with a shared **consensual** clustering **z**.

$$
\begin{aligned}
\pi &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
z_i &\sim \text{Discrete}(\boldsymbol{\pi}) \\
\theta_j^{(f)} &\sim G_0^{(f)} \\
\theta_j^{(a)} &\sim G_0^{(a)} \\
a_u | z_i, \theta_{z_u}^{(a)} &\sim F^{(a)}(\theta_{z_u}^{a}) \\
f_u | z_i, \theta_{z_u}^{(f)} &\sim F^{(f)}(\theta_{z_u}^{f}) \\
y_u &\sim g(f_u)
\end{aligned}
$$



- Users in the same cluster have similar features **a** and behaviors **f**.
- If not enough data to infer latent f_u, leverage data from users in the same cluster.

(Potentially) Infinite clusters

Chinese Restaurant Process

We assume a *Chinese Restaurant Process* prior (a form of Dirichlet Process) on the cluster assignments. That is, we assume that users choose their cluster one by one with probabilities:

$$p(z_u = j|\mathbf{z_{-u}}) \propto n_j \qquad \text{if not empty}$$
$$p(z_u = j|\mathbf{z_{-u}}) \propto \alpha \qquad \text{if empty}$$

where n_j is the number of users already in cluster j.

This makes the model *non-parametric in the number of clusters*: the number of clusters K is also inferred from the data.

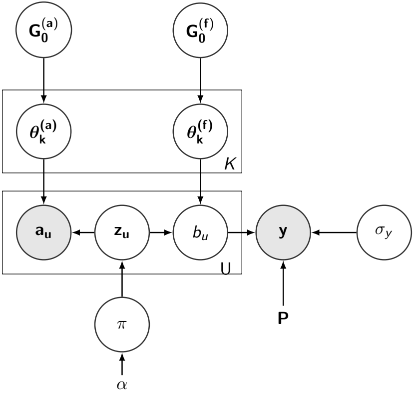
Hypothetical scenario

Behavior = catalytic power for thread length

- Each user has catalytic power b.
- The final length of a thread $\mathbf{y}_i$ is the sum of catalytic powers of the first N users.

$$\mathbf{y} \sim \mathcal{N}(\mathbf{P}^{\mathbf{T}}\mathbf{b}, \sigma_y \mathbf{I})$$

  P: binary participation matrix. $p_{ut} = 1$ if user u is among the first participants of thread t.
- User features $\mathbf{a}_u$ and latent coefficients b_u drawn from mixture of Gaussians.



The more threads/user we have, the easier to learn coefficients **b**.

Inference

MCMC

We chose a **Gaussian mixture model** for both views, following the Infinite Gaussian Mixture Model[8][9].

- **Gibbs Sampling** for most of the variables
- Except for degrees of freedom of Wishart distributions, sampled by **Adaptive Rejection Sampling**.
- 30,000 samples of each variable, the first 15,000 dropped-out (burning).

[8] Carl E Rasmussen. "The infinite Gaussian mixture model". In: *Advances in Neural Information Processing Systems 12*. Ed. by S A Solla, T K Leen, and K Müller. Cambridge, MA: MIT Press, 2000, pp. 554–560.

[9] Dilan Görür and Carl Edward Rasmussen. "Dirichlet process Gaussian mixture models: choice of the base distribution". In: *Journal of Computer Science and Technology* 25.July (2010), pp. 653–664.

Benchmark

Compared models:

- `dual-DP`: dual-model with infinite clusters
- `dual-fixed`: dual-model that knows the number of clusters
- `single`: model with no clusters (only learns latent coefficients)

Metrics:

- Predictions: (negative) loglikelihood of test set:

$$p(\mathbf{y}^{(test)}|\mathbf{y}^{(train)})$$

- Clustering: Adjusted Rand Index (ARI)
    - Measures pairwise discrepancy with true cluster.

Data

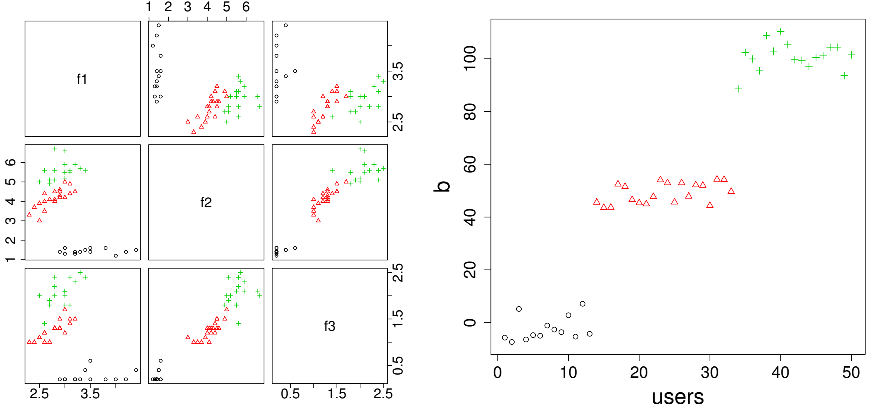


Figure : User features (left) and user *latent* coefficients (right)

Results

Dual-view models learn with less data (less examples per user).

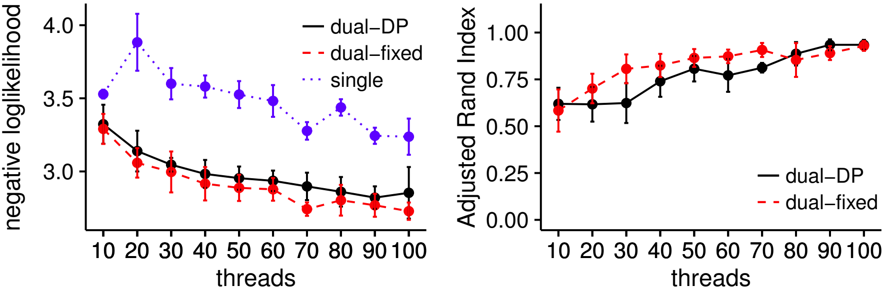


Figure : Results for the iris dataset. Comparison of models under different threads/users ratios (50 users and variable number of threads). Means and standard errors over 5 runs.

Discussion

- Dual-view models learn more with less.
- Warning: the model looks for consensus, do not use contradictory information between the views!
- Gibbs inference very slow for large data.

Possible improvements over inference:

- Easier inference: one group, one behavior.
- Variational Bayes for large scale inference.

Outline

1. Introduction and data
2. Role detection based on conversations motifs
3. Role detection based on behavioral functions
4. Role detection based on features and behavioral functions
5. Conclusions

General conclusions

Contributions

Conversation-based roles:

- Detection of roles based on conversation structures.
  - Order-based neighbourhood can detect **different types of conversationalists not detectable by non-structural methods** (initiators, terminators, root repliers, debaters,...)
- Behavior function to model, detect and predict different types of behaviors.
  - **Extreme roles are predictable.**
- A dual-view model to integrate features and functional/behavioral data.
  - **Learns more with less data**.

General conclusions

Perspectives

Adapting the dual-view model:

- *Features*: Motif attributes (from a Discrete distribution instead of Normal).
- *Behaviors*: $p(\pi_t \sim i) \propto \alpha d_i + \beta r_i + \tau^{l_i}$ lacks a conjugate prior. Gibbs sample not possible. Instead: M-H, MAP,...

Structure + Language:

- Language may provide useful information (sentiment, topic, type of content: help, discussion, (dis)agreement,...)

Roles or not roles?:

- Conjecture: Forums have some sets of users with clear behavioral roles, and a majority with no specific role (variable behavior).
- Need of *predictive tests* to confirm that roles are roles and not just collections of different types of past behaviors.

Publications

Lumbreras A., Guégan M., Velcin J., Jouve B. (2016) Non-parametric clustering over user features and latent behavioral functions with dual-view mixture models. *Computational Statistics*.

Lumbreras A., Guégan M., Julien J., and Jouve B. (2015) Clustering users features and latent behavioral functions. *In StatLearn* [Poster]

Lumbreras A., Lanagan J., Velcin J., Jouve B. (2013). Analyse des rôles dans les communautés virtuelles : définitions et premières expérimentations sur IMDb. *Modèles et Analyses Réseau : Approches Mathématiques et Informatiques (MARAMI)*

Lumbreras A., Lanagan J, Jouve B., Velcin J. (2013). An insight into the Analysis of Roles in IMDb. *Workshop on Complexity in social systems: from data to models*, Cergy Pontoise (95), 27-28 juin 2013

Merci