manuscript No.
(will be inserted by the editor)

You are the way you structurally talk: Structural-temporal neighborhoods of posts to characterize users in online forums

Alberto Lumbreras · Julien Velcin ·
Marie Guégan · Bertrand Jouve

Draft date: March 29, 2016

1 Introduction

The popularization of online forums has brought a growing interest on their underlying dynamics. As any other complex system, an the dynamic of online forums can be studied at different levels, from the more macro to the most micro. Macro dynamics are, for instance, the evolution of some global properties such as its diameter, or its distribution degree. Micro dynamics are, for instance, the triadic motifs that represent local phenomena such as transitivity (friends of my friends are also my friends).

An interesting question in online communities is that concerning roles. In sociology, roles are generally seen as the set of expected behaviours that are attached to a position in the community. Extrapoliting the notion of role, some researchers have looked for roles in online forums. Some others have tried to detect the roles and the users who hold that roles.

Roles can also be studied from the macro or the micro perspective. If studied from the macro, we can analyse the number of users, the percentage of replied post, its centrality in the network, and so forth.

Alberto Lumbreras · Marie Guégan
Technicolor
975 Avenue des Champs Blancs,
35576 Cesson-Sévigné,
France
E-mail: alberto.lumbreras@technicolor.com
E-mail: marie.guegan@technicolor.com

Julien Velcin
Laboratoire ERIC, Université de Lyon,
5, avenue Pierre Mendès France, 69676 Bron,
France
E-mail: julien.velcin@univ-lyon2.fr

Bertrand Jouve
Université de Toulouse; UT2; FRAMESPA/IMT; 5 allée Antonio Machado, 31058 Toulouse, cedex 9
CNRS; FRAMESPA; F-31000 Toulouse
CNRS; IMT; F-31000 Toulouse
France
E-mail: jouve@univ-tlse2.fr

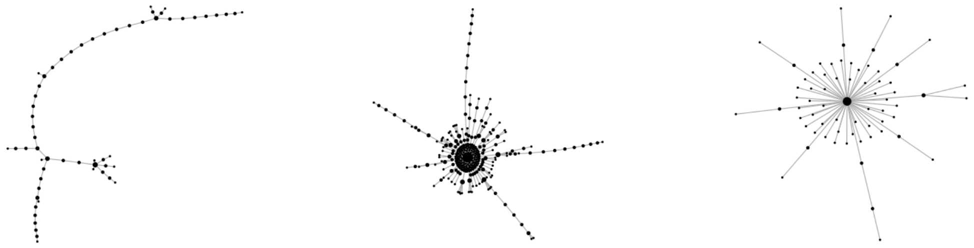

Fig. 1 Representation of discussion threads with post tree graphs. Vertex represent posts and edges represent replies between posts.

In this paper, we focus on the analyse of roles at a micro level, and more specifically at the discussion level. We would like to answer the following question: are there different types of users in terms of the kind of conversation they participate in?

Our intuition is that some users like participating in some kind of discussion rather than other. Certainly, an analysis of the textual content will tell us much about a discussion. However, due to the huge diversity of topics, vocabulary, and the difficulty of current algorithms to capture the language subtleties such as humor, irony, or changing contexts, we turn our attention towards the structure of the discussions. More precisely, we analyze the local graph structure in which a user post in embedded, in the hope that this structure will be meaningful since it also reflects the kind of conversation in that part of the thread.

Formally, these local graphs are known as neighborhoods.

The remaining of the paper is as follows. First we will introduce our representation of online discussions. Then we will explain our concept of structural neighbourhood. Then we will apply this to the clustering of users and we will analyze the clusters. Finally, we will attemp to study whether these roles have predictive power.

2 Discussion trees

A natural way to represent a conversation thread is a tree graph where vertex represent posts and edges represent replies from some post to another. The tree is rooted at the post who started the thread. This representation allows to apply the mathematical toolbox of graph theory. Figure 1 shows some real examples of trees in a Reddit[1] forum.

3 Neighborhoods in discussion trees

Our goal is to characterize conversations by their local structure. By local structure we mean the structure of the graph that surrounds a given post, which we call *neighbourhood*. Given a set of posts written by a user, analysing the neighbourhoods where user posts are embedded will give us an insight of what kind of user she is. It is important to note that, if the space of possible neighbourhoods is too small, we might not be able to say anything interesting, while if it is too large we might have neighbourhoods that correspond to the same type of conversation.

[1] www.reddit.com

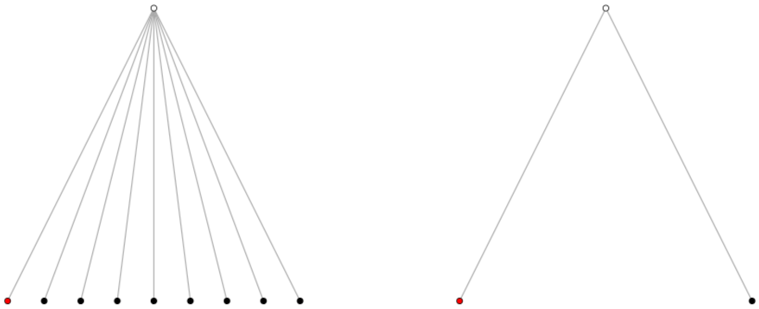

Fig. 2 The size of the neighbourhood with radius r is unbounded. These two graph represent frequent neighbourhoods of a post (red) that replied to the root (white). However, the tree in the left corresponds to a very successful root while the three in the right has not brought the attention of too many users.

In this section, we discuss about the convenience of the classic neighbourhood definition in dynamic graphs such as discussion trees. Then we introduce a new definition to overcome some of the limitations of the classic one when dealing with temporal trees.

3.1 Structural neighborhoods in discussion trees

Definition 1 Given a tree graph G, the *structural neighbourhood* of radius r of post i, denoted as $\mathcal{N}_i(r)$, is the induced graph composed of all the vertices that are at distance equal or less than d from post i.

The above definition has two drawbacks when used in the context of conversation trees. First, the dynamics of the conversation (time or order in which posts are attached to the tree) are not entirely captured in the structure of a tree representation. We know, for instance, that the time in which a node was attached to the tree is always posterior to that of its parent. But it impossible to say, by just looking at the structure of the tree, the order in which a set of sibling posts replied to their common parent post. Since these siblings may get more replies themselves, a single *structural neighbourhood* may sometimes correspond to very different dynamics (see Figure). Second, a structural neighbourhood at a given radius r has an unbounded number of posts, and therefore the space of possible neighbourhoods is infinite. This poses a problem when trying to categorize conversations since many conversations, while structurally different, can be considered semantically equivalent (see Figure)

3.2 Structuro-temporal neighborhoods in discussion trees

Definition 2 Given a tree graph G, the *structural-temporal neighborhood* of distance d of post i, denoted as $\mathcal{N}_i^T(r,n)$ is the induced subgraph from its structural neighborhood composed the n vertices that are closer to i in time and for which there exists a path to i in $\mathcal{N}_i^T(r,n)$.

The above definition has two advantages over the previous one. First, dynamics are better taken into account since the neighbourhood only includes posts that are structurally and temporally close. Second, the size of the neighborhood has an upper bound of $max(\mathcal{N}_i(r)), o)$.

Note that this definition does not consider the timestamp of posts but only the order in which they arrived. An alternative definition of structural-temporal neighborhood might include not the o posts that are closest to the ego post, but every post j at distance $|t_i - t_j| < \tau$ from the ego post i. However, both options consider that the local dynamic ends at a fixed horizon (o or τ). This is specially problematic for the case of τ since different conversation threads may grow at different speeds, as well as different parts of the same thread or even the same part of the thread at different moments (e.g. threads barely grow at night). If we were to use τ as temporal horizon, we need a mechanism to automatically decide the proper τ for every post.

Definition 3 Let $s = p^{(1)}, ...p^{(m)}$ be the sequence of posts in the structural neighbourhood $\mathcal{N}_i(r)$ ordered by their temporal distance to i. Let $\pi = \{\Delta t_1, ..., \Delta t_M\}$ be the sequence of time differences between a post in s and its predecessor in s. The *temporal radius* τ of the neighbourhood is the timestamp of the last post k whose Δt_k comes from the same distribution than its predecessors $\Delta t_{k-1}, ..., \Delta t_1$ (an intuition for this definition is given in Figure 3)

According to the above definition, to find the temporal radius we need to find the first outlier in a sequence of Δt. Assuming the Δt_i comes from an Exponential distribution, and that its λ parameter comes itself from a Gamma distribution, we can exploit this conjugacy and obtain the analytical form of the probability of a given Δt_k of belonging to the same distribution than its predecessors:

$$p(\Delta t_k|\Delta t_1, ...\Delta t_{k-1}) = \int p(\Delta t_k|\lambda_{k-1})p(\lambda_{k-1}|\Delta t_1, ...\Delta t_{k-1}) = \frac{(\beta + n\overline{x})^{\alpha+n}}{(\beta + n\overline{x} + \Delta t_k)^{\alpha+n+1}}(\alpha + n) \tag{1}$$

Details of this derivation are given in the Appendix.

3.3 Colored neighborhoods

In conversation trees, the root vertex plays an important role since it represents the post that opened the conversation and thus set the topic of the thread. Therefore, a reply to the root post is clearly different from a reply to any other post in the tree. We distinguish root posts from non-root posts by assigning root posts a unique color (white). We also assign a unique color to the ego post (red).

4 Application to Reddit forums

In this section, we use our definition of structural-temporal neighbourhood to detect different types of users according to the neighbourhoods in which their posts are embedded.

4.1 Podemos forum

Our dataset consists of all posts from March 2014 to May 2015 of the Podemos forum in Reddit[2][3]. The Podemos forum was conceived in March 2014 as a tool for internal democracy, and forum

[2] https://www.reddit.com/r/podemos

[3] I have also the data for *gameofthrones, france, datascience, machinelearning complexsystems, philosophy, twoxchromosomes, trees, sex*

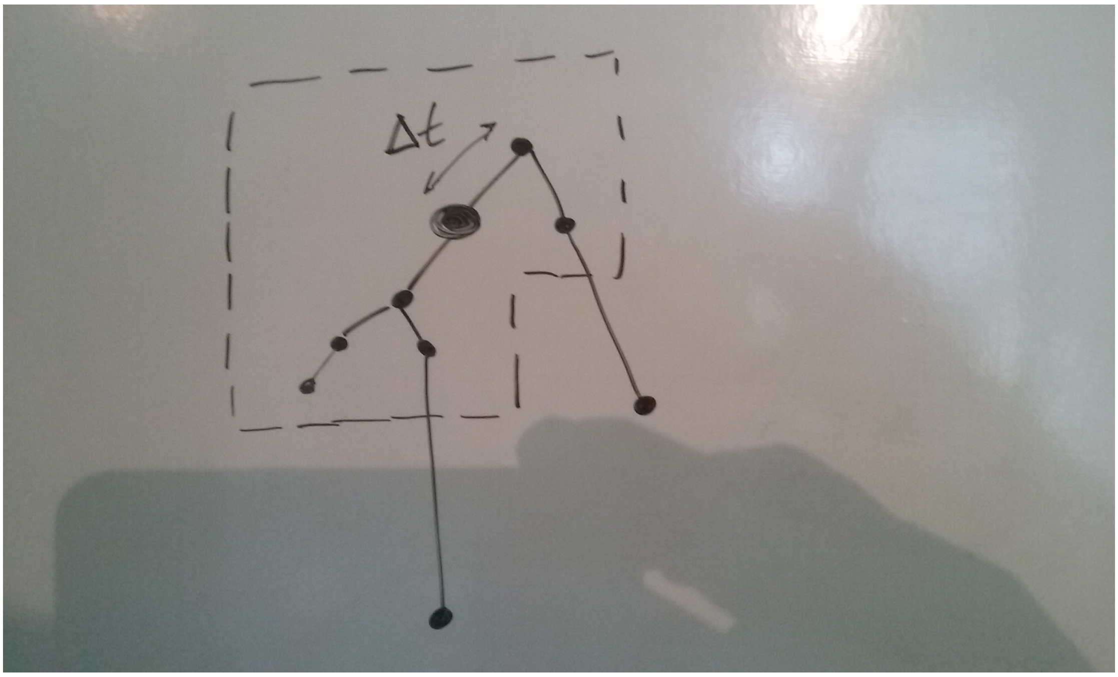

Fig. 3 Intuition to choice the temporal cutpoint of a neighbourhood. The two posts out of the box have a distance (time) to their parents clearly bigger than the average, so we consider that they are not part of the same local dynamic.

members used it to debate ideological and organizational principles that were later formalized in their first party congress hold in Madrid the October 18th and 19th 2014. Nowadays, its members use it mainly to share and discuss about political news. The forum contains 83,6119 posts spread over 47,803 threads and written by 26,193 users. (see Figure 4)

(Il fait que j'implemente l'équation 1 pour chercher les neighborhoods. Puis il faut voir si elle sert tout à fait à bien detecter la baisse du rythme)

4.2 Counting neighorhoods

In this section we will build a matrix M for one of the forums of Reddit and then we will cluster users to find groups that have similar preferences over the different types of conversations represented by the neighbourhood structures.

If we set the a spatial radius $r = 1$ and an order $n = 4$, the temporal neighborhood $N_G^{(t)}(i, 1, 4)$ recovers the four adjacent posts that were written temporally closest to a given post. Figure 5 shows the four detected neighborhoods and their frequency.

In order to detect more complex types of neighborhoods we enlarge the spatial radius to $r = 4$. Figure 6 shows the eight detected neighbors and their frequency. It is interesting to see different structures like chains or stars where the root node might or might no appear and the ego post is placed at different positions.

Note that some neighbors are still very similar. For instance, *common answer 3* is very close to *common answer 2* in the sense that they both represent the ego post replying to a root post

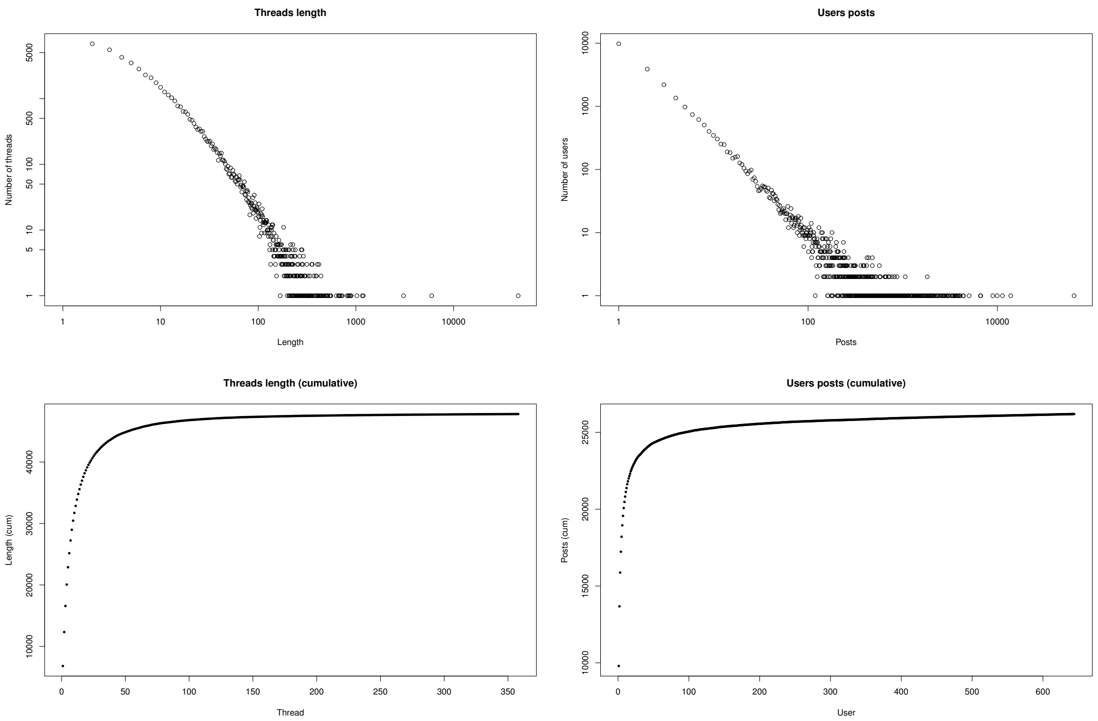

Fig. 4 Log-log and cumulative distributions and of thread lengths and posts per user.

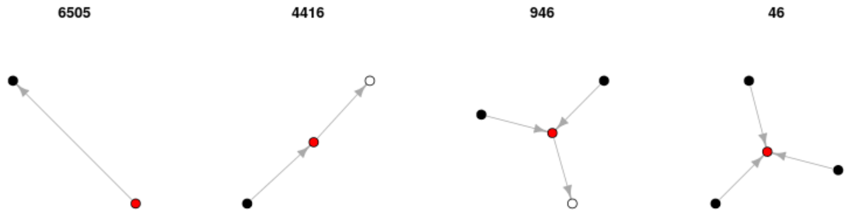

Fig. 5 Neighborhood counts with $r = 1$ and $n = 4$

that has one and two replies respectively. It seems a good idea to merge them both since they seem to represent the same type of discussion.

4.3 Clustering

Given the neighborhoods in which each user participates, we will analyze whether there exists different types of users or all users look similar.

We want to make the analyze independent of the number of posts, and for that we normalize each user feature vector so that features indicate the percentage of posts in this kind of neighborhood. Moreover, some neighborhoods are much more common than others due to the nature of the forums. Thus, we normalize and scale the features so that every feature has a global mean

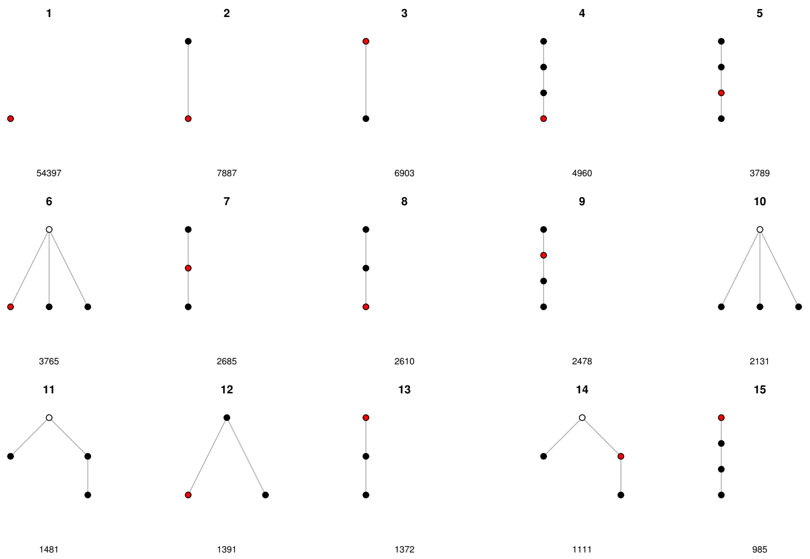

Fig. 6 Most frequent neighborhoods with $r = 4$ and $n = 4$

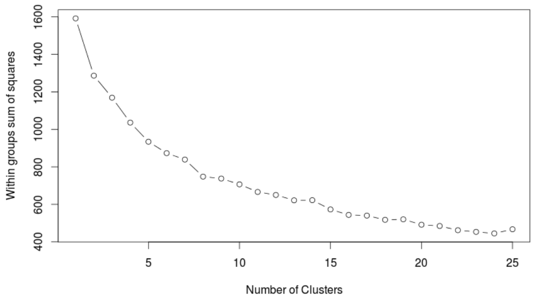

Fig. 7 Within-Sum of Squares under k-means

0 and variance 1. User features now represent z-scores, that is, how many standard deviations is this user feature away from the mean.

We use a simple k-means to find the clusters. To decide the number of clusters, we run k-means for k=2,...,25 clusters and look at the Within-Cluster Sum of Squares (Figure 7) and we chose $k = 5$ so that the results are more interpretable. Figure 8 shows a PCA projection of the users colored by cluster and the distribution of the clusters in every dimension.

4.4 Predictive roles

One interesting use of roles is to make predictions on users behaviors. Here, we will analyze whether the initial composition of roles in a thread is a good predictor of the final length of the thread.

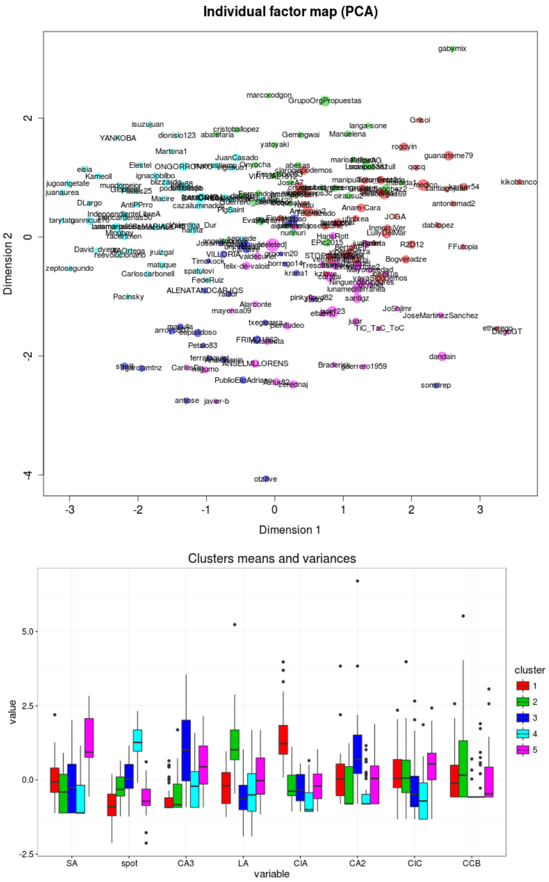

Fig. 8 PCA projection of the clusters found and cluster profile in every dimension.

5 Conclusions

Appendix

$$p(\Delta t_k|\Delta t_1,...\Delta t_{k-1}) = \int p(\Delta t_k|\lambda_{k-1})p(\lambda_{k-1}|\Delta t_1,...\Delta t_{k-1}) \tag{2}$$

$$= \frac{(\beta + n\overline{x})^{\alpha+n}}{(\beta + n\overline{x} + \Delta t_k)^{\alpha+n+1}}(\alpha + n) \tag{3}$$

First we work on the posterior distribution over λ_{k-1}:

$$p(\lambda_{k-1}|\Delta t_1,...\Delta t_{k-1}) \propto p(\Delta t_1,...\Delta t_{k-1}|\lambda_{k-1}|)p(\lambda_{k-1}) \tag{4}$$

$$\propto \prod \lambda e^{-\lambda x}\lambda^{\alpha-1}e^{-\beta\lambda} \tag{5}$$

$$\propto \lambda^{\alpha+k-1}e^{-(\beta+k\overline{x})\lambda} \tag{6}$$

$$\propto \mathcal{G}(\alpha + k, \beta + n\overline{x}) \tag{7}$$

And then the predictive posterior is:

$$p(\Delta t_k|\Delta t_1,...\Delta t_{k-1}) = \int p(\Delta t_k|\lambda_{k-1})\mathcal{G}(\lambda_{k-1}|\alpha + k, \beta + n\overline{x}) \tag{8}$$

$$= \frac{(\beta + n\overline{x})^{\alpha+k}}{\Gamma(\alpha + k)}\int \lambda e^{-\lambda\Delta_k}\lambda^{\alpha+k-1}e^{-(\beta+n\overline{x})\lambda}\mathrm{d}\lambda \tag{9}$$

$$= \frac{(\beta + n\overline{x})^{\alpha+k}}{\Gamma(\alpha + k)}\int \lambda^{(\alpha+k+1)-1}e^{-(\beta+n\overline{x}+\Delta_k)\lambda}\mathrm{d}\lambda \tag{10}$$

$$= \frac{(\beta + n\overline{x})^{\alpha+k}}{\Gamma(\alpha + k)}\frac{\Gamma(\alpha + k + 1)}{(\beta + n\overline{x} + \Delta_k)^{\alpha+k+1}} \tag{11}$$

$$= \frac{(\beta + n\overline{x})^{\alpha+k}}{(\beta + n\overline{x} + \Delta_k)^{\alpha+k+1}}(\alpha + k) \tag{12}$$