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

Alberto Lumbreras Julien Velcin
Marie Guégan Bertrand Jouve

I. Introduction

The popularization of online forums has brought a growing interest on their underlying dynamics. As any other complex system, the dynamic of online forums can be studied at different levels, from the most macro to the most micro. Macro dynamics are, for instance, the evolution of some global properties of the social graph 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. Extrapolating 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.

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 humour, irony, or changing contexts, we turn our attention towards the structure of the discussions. More precisely, we analyse 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 neighbourhoods.

The remaining of the paper is as follows. We first discuss about the convenience of the classic neighbourhood definition in dynamic graphs such as discussion trees. Then we introduce two new definitions that takes into account the time to overcome some of the limitations of the structural neighbourhood. We will finally apply these time-based neighbourhoods to detect clusters of users than tend to appear in the same time-based neighbourhoods.

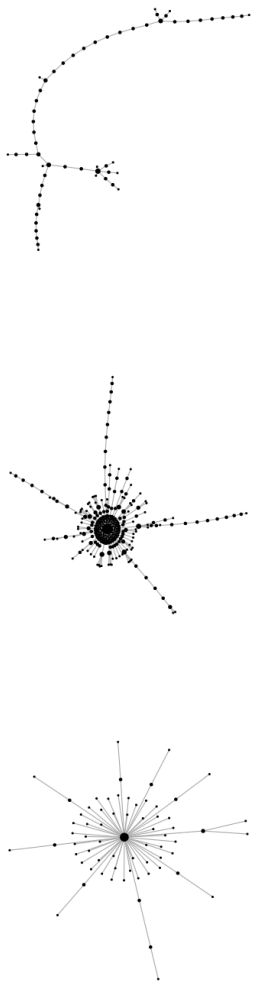

Fig. 1. Representation of discussion threads with post tree graphs. Vertex represent posts and edges represent replies between posts.

II. Discussion trees

We represent a conversation thread as 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. Figure 1 shows some real examples of trees in a Reddit[1] forum.

The *root* post of the tree is the post that starts the discussion. It is the only post that has no parent. A *leaf* is a post with no replies. A *branch* of the tree is the shortest path between the root and some of the leafs. Two branches may share their first posts.

III. Structural neighbourhoods in discussion trees

Extending the classic definition of neighbourhood according to which two vertices are neighbours if the distance between them is one, in this paper we define the *structural neighbourhood* as follows:

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.

This 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. Thus, a single *structural neighbourhood* may sometimes correspond to very different dynamics. 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 2).

IV. Structuro-temporal neighbourhoods in discussion trees

In this section, we propose two new time-based definition of neighbourhoods.

[1] www.reddit.com

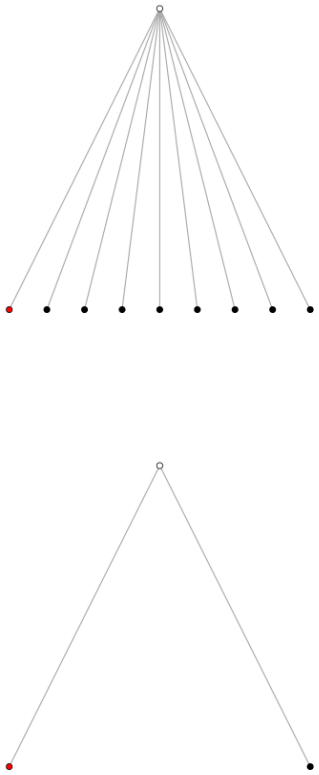

Fig. 2. The size of the neighbourhood with radius r is unbounded. These two graph represent frequent neighbourhoods of a post (red) that replied to the root (white). However, the tree in the left corresponds to a very successful root while the three in the right has not brought the attention of too many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

A. Order-based

Our first definition is based on the order in which posts are attached to the tree.

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

This definition has two advantages over the *structural neighbourhood*. First, the temporal aspect of the conversation is better taken into account since the neighbourhood

only includes posts that are structurally and temporally close. Second, the size of the neighbourhood has an upper bound of $max(\mathcal{N}_i(r)), o)$.

B. Time-based

Our second definition uses the real timestamps of posts in order to decide where the boundaries of a neighbourhood are. We might naively set fixed time-based boundaries for the neighbourhood by only including in the neighbourhood those posts whose timestamp p_i is at distance less than τ from the ego post $|t_i - t_{ego}| < \tau$. However, the pace at which posts are added to the conversation may be very different between conversations, and even between different parts of the same conversation tree. Instead, we propose a definition based on the concept of *local dynamic*.

Definition 3: The temporal neighbourhood $\mathcal{N}_i^T(r)$ is the maximal subgraph of the structural neighbourhood $\mathcal{N}_i(r)$ where all the posts belong to the same local dynamic than the ego post.

Our definition of *local dynamic* is based on the detection of changepoints in the time stamps of the posts. Given a sequence of consecutive posts in the same branch b of a tree, denoted as $p_s, ..., p_e$ we say that they belong to the same vertical dynamic if there is no (vertical) changepoint p_i in b such that $p_1 \prec p_i \preceq p_n$. Similarly, given a sequence of chronologically sorted siblings s, denoted as $p_s, ..., p_e$, we say that they belong to the same horizontal dynamic if there is no (horizontal) changepoint p_i in s such that $p_1 \prec p_i \preceq p_n$. Now we can outline the complete algorithm for time-based neighbourhood (Algorithm 1). Figure 3 illustrate a tree with a set of breakpoints and the temporal neighbourhood for a given node.

C. Colored neighbourhoods

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).

V. 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.

Data: Posts tree g, vertical breakpoints, horizontal breakpoints, ego post ego
Result: Subgraph of g with all vertices in $V(g)$
Compute structural neighbourhood $\mathcal{N}_i(r)$;
ancestors $\leftarrow$ ancestors(ego) in $\mathcal{N}_i(r)$;
older_siblings $\leftarrow$ older_siblings(ego) in $\mathcal{N}_i(r)$;
dump_posts $\leftarrow \varnothing$;
for *bp* $\in$ *vertical breakpoints* **do**
  if *bp* $\in$ *ancestors* **then**
    dump_posts $\leftarrow$ dump_posts $\cup$ ancestors(bp);
  else
    dump_posts $\leftarrow$ dump_posts $\cup$ descendants(bp) $\cup$ bp;
  end
end
for *bp* $\in$ *horizontal breakpoints* **do**
  if *bp* $\in$ *(older_siblings* $\cup$ *ancestors)* **then**
    dump_posts $\leftarrow$ dump_posts $\cup$ older_siblings(bp);
  else
    dump_posts $\leftarrow$ dump_posts $\cup$ younger_siblings(bp) $\cup$ bp;
  end
end
$\mathcal{N}_i^{(t)}(r) \leftarrow$ delete(dump_posts) from $\mathcal{N}_i(r)$;
Algorithm 1: Extraction of time-based neighbourhood

A. 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 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)

B. 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

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

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

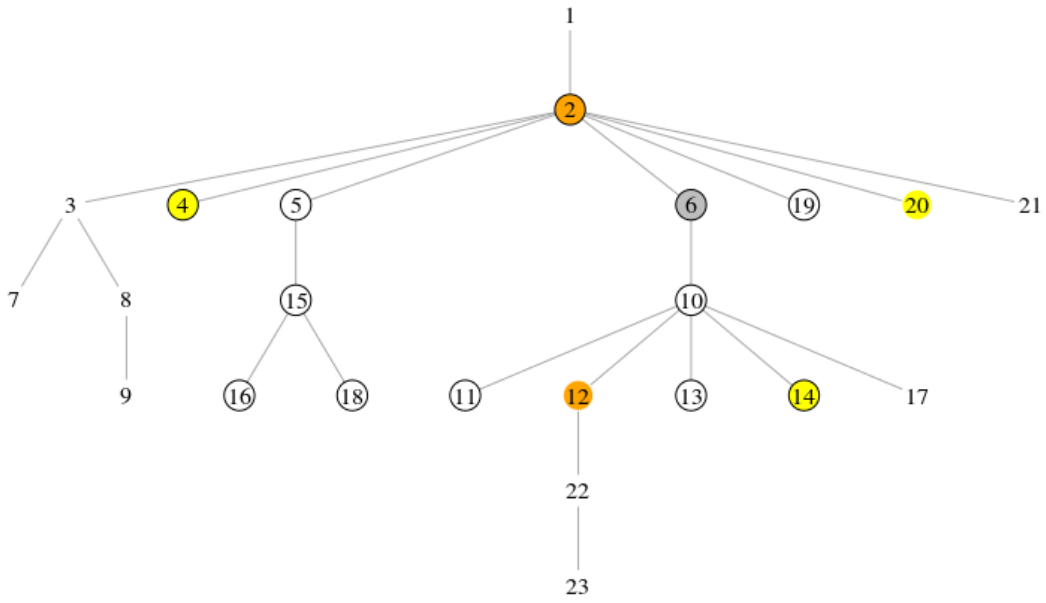

Fig. 3. Time-based neighbourhood. Horizontal changepoints (yellow) and vertical changepoints (orange) represent posts that are temporally far from their predecessors (siblings or parents) and therefore set the limits of the neighbourhood. Post 6 is the ego, and posts with circular borders are those belonging to the time-based neighbourhood.

types of conversations represented by the neighbourhood structures.

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

In order to detect more complex types of neighbourhoods 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 neighbours 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 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.

C. Clustering

Given the neighbourhoods 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

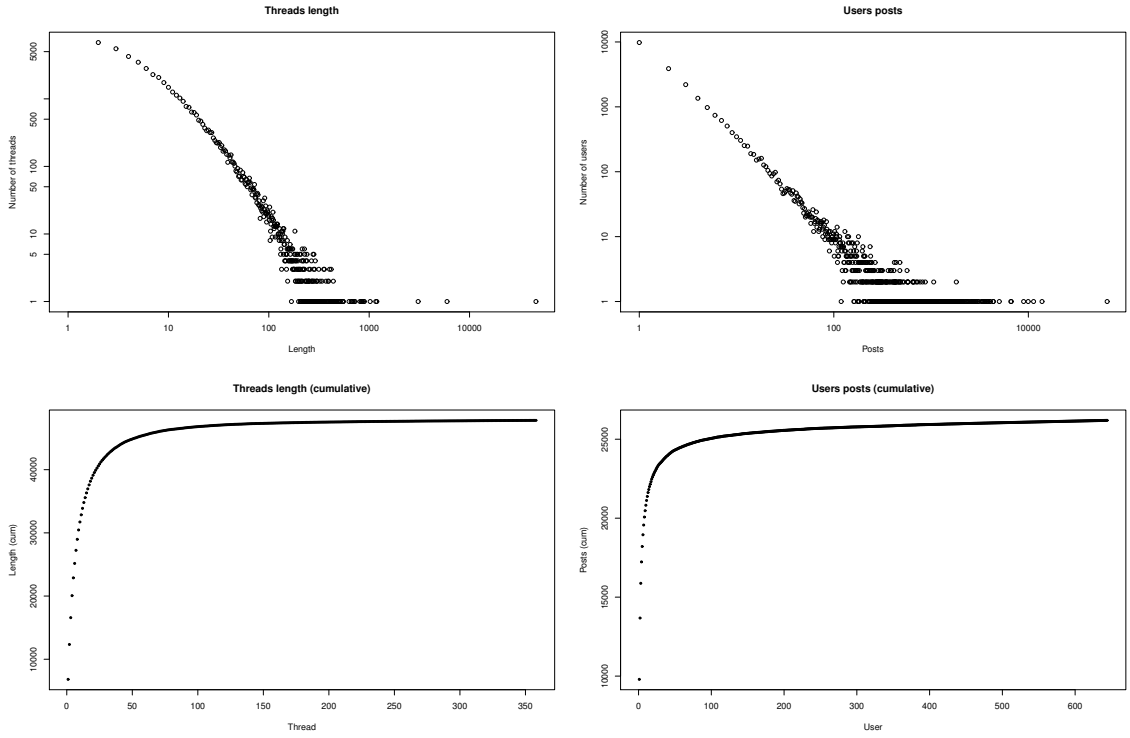

Fig. 4. Log-log and cumulative distributions and of thread lengths and posts per user.

feature vector so that features indicate the percentage of posts in this kind of neighbourhood. Moreover, some neighbourhoods 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 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 **??**) and we chose $k = 5$ so that the results are more interpretable. Figure 7 shows a PCA projection of the users colored by cluster and the distribution of the clusters in every dimension.

VI. Conclusions

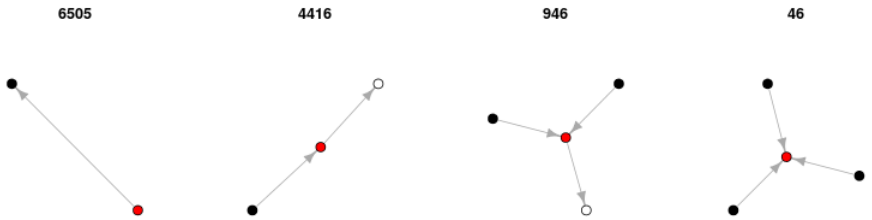

Fig. 5. neighbourhood counts with $r = 1$ and $n = 4$

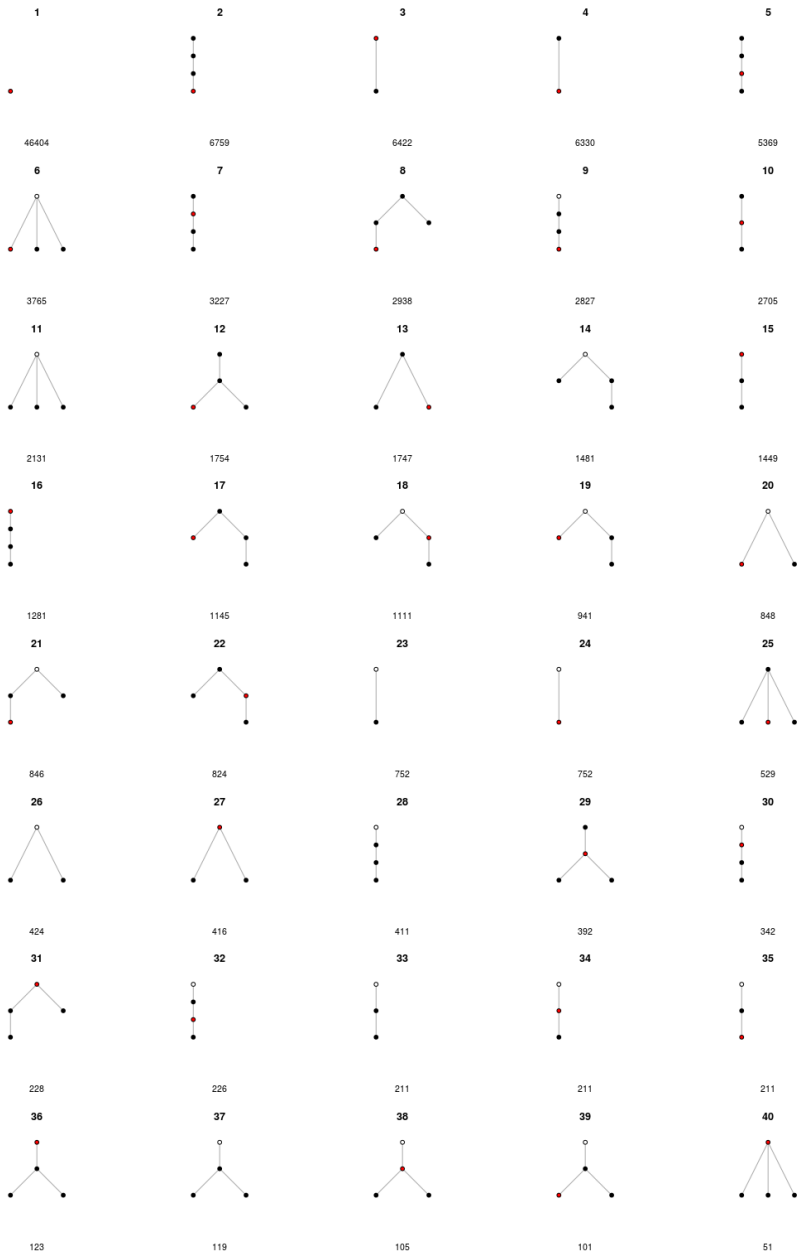

Fig. 6. Order-based neighbourhoods with $r = 3$ and $n = 4$

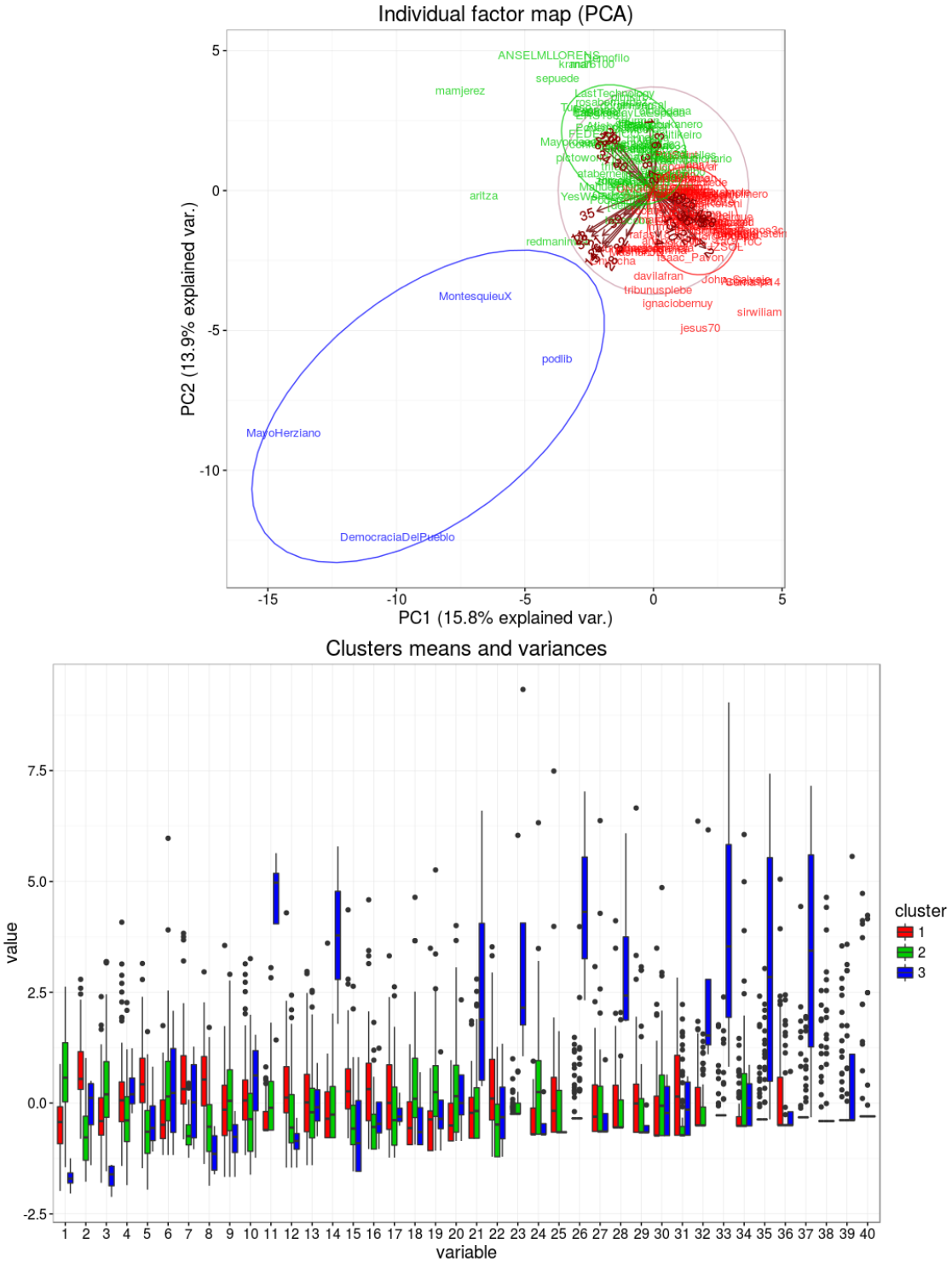

Fig. 7. PCA projection of the clusters found and cluster profile in every dimension.