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

(Draft version 13/04/2016)

Alberto Lumbreras*, Bertrand Jouve†, Marie Guégan‡, Julien Velcin§
*Technicolor, France
Email: alberto.lumbreras@gmail.com
†Université de Toulouse; UT2; FRAMESPA/IMT,
5 allée Antonio Machado, 31058 Toulouse, cedex 9
Email: jouve@univ-tlse2.fr
‡Technicolor
975 Avenue des Champs Blancs
35576 Cesson-Sevigné,
France
Email: marie.guegan@technicolor.com
§Laboratoire ERIC, Université de Lyon,
5 avenue Pierre Mendès France, 69676, Bron
France
Email: julien.velcin@univ-lyon2.fr

***Abstract*—Users of social networks are often characterised by extracting some relevant features from the graphs associated to that network. The structural dynamic of the conversation are usually forgotten due to a lack of proper tools. We present a purely graph-based method to cluster users in online forums.**

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) and that may explain some macro phenomena.

The study of user behaviours has also brought the attention of researchers. 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 present two types of structural patterns adapted to discussion trees. These trees can represent e-mails, forums, etc. Here we will use forum data to illustrate out methodology.

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 richer definitions that take into account the order and the time. To illustrate the kind of neighbourhoods we obtain in a real forum, we show the neighbourhood census in a Reddit forum. Finally, we apply both neighbourhood definitions to detect clusters of users that tend to appear in the same

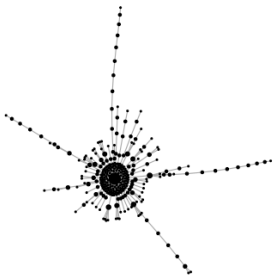


Fig. 1. Representation of discussion threads with post tree graphs. Vertex represent posts and edges represent replies between posts.

neighbourhoods. We close the paper with some suggestions of future research.

II. Discussion trees

The interactions like those found in an online forum can be modelled, at least, by two sort of graphs. On the one hand, we can create a graph where vertices represent individuals and an edges represent an interaction between two individuals. At the individual level, this graph is suitable for a large set of analysis concerning the centrality of users, the detection of communities, the detection of groups that tend to interact with the same other groups. At the community level we can analyse properties such as the density of the graph, the clustering coefficient or the degree distribution. On the other hand, we can create a separate tree graph for every discussion where posts are represented by the vertices and and edge from one vertex to another indicates that first post is a reply to the second. This allows to analyse things like patterns of time delay between posts [1].

We can even mix different levels of representations to study the relationship between the properties and dynamics among the different levels [2].

Of course, we are not limited to the study of graphs and some studies take into account the content of the conversations, the cross-posting activity of users, and so forth [3]

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.

Others: [4]

III. Structural neighbourhoods

Extending the classic definition of neighbourhood according to which two vertices are neighbours if the distance between them is one, we start by the following definition of *structural neighbourhood*:

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.

In the context of discussion threads, this definition has two limitations. First,

the decision on whether to include some post in the neighbourhood is only based on the structural distance, and therefore two posts that are at distance $d \leq r$ are considered neighbours of i regardless of the time they were written. In conversations, time plays an important role, therefore this is an important limitation. Another consequence of looking only at the structure is that the number of possible neighbourhoods within a radius r is infinite. This poses a problem when trying to categorize conversations since many conversations, while structurally different, can be considered semantically equivalent.

IV. Structuro-temporal neighbourhoods

As we have discussed, structural neighbourhood falls short in the analysis of conversational structures. Instead, we propose two new definitions of neighbourhood that take into account order and time in which posts are written.

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 $max(\mathcal{N}_i(r)), n)$ thus making the space of possible neighbourhood structures finite.

B. Time-based

Our second definition uses the real timestamps of posts in order to automatically decide where the boundaries of a given 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, making it difficult to set a fixed temporal radius τ. Instead, we propose a definition based on the concept of *local dynamic*.

Our definition of *local dynamic* is based on the detection of changepoints in the time stamps of the posts (ref). 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

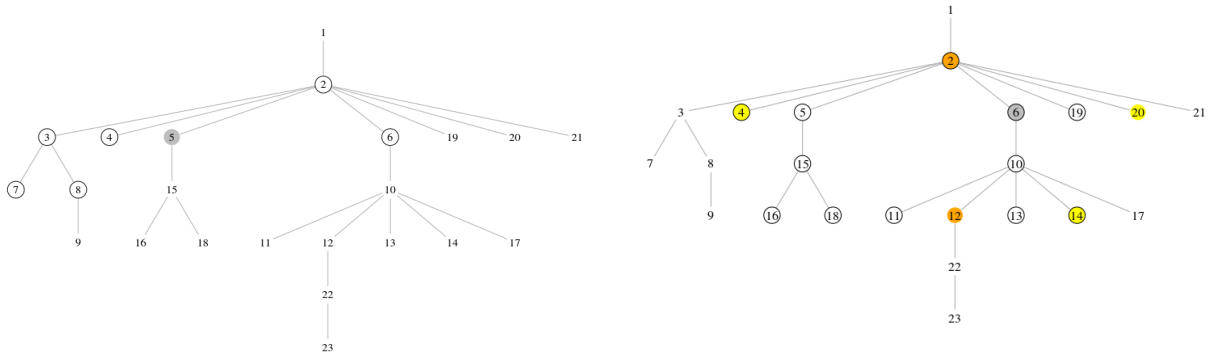


Fig. 2. Illustration of order-based (left) and time-based (right) neighbourhoods. Grey nodes represent the ego post. Nodes with circles are those in the neighbourhood. Numbers indicate the order of the post. The order-based neighbourhood has parameters $r = 3$, $n = 6$. In the time-based neighbourhood, 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.

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 $p_1 \prec p_i \preceq p_n$.

Definition 3: Given an ordered tree T, 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.

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.

Algorithm 1 Extraction of time-based neighbourhood

Input: Posts tree g, vertical breakpoints, horizontal breakpoints, ego post ego
Output: Subgraph of g with all vertices in $V(g)$

```
Compute structural neighbourhood N_i(r)
ancestors ← ancestors(ego) in N_i(r)
older_siblings ← older_siblings(ego) in N_i(r)
dump ← ∅
for bp ∈ vertical breakpoints do
  if bp ∈ ancestors then
    dump ← dump ∪ ancestors(bp)
  else
    dump ← dump ∪ descendants(bp) ∪ bp
  end if
end for
for bp ∈ horizontal breakpoints do
  if bp ∈ (older_siblings ∪ ancestors) then
    dump ← dump ∪ older_siblings(bp)
  else
    dump ← dump ∪ younger_siblings(bp) ∪ bp
  end if
end for
N_i^(t)(r) ← delete(dump_posts) from N_i(r)
```

C. Colored neighbourhoods

Once we extract a neighbourhood, we assign colors to their nodes to represent some relevant characteristics. In particular, we assign the red color to the ego post, the orange colour to all other posts written by the author of the ego (to identify re-entries) a white color to the root (if included in the neighbourhood), the yellow color to all posts written by the same author who wrote the parent of ego (to identify, for instances, debates between the ego author and the other author), the grey color to all posts written

before the ego and the black color to all posts written after the ego. Since some posts might correspond to several colors, we perform a sequential assignment of colors given by `grey, black, orange, ego, yellow, white`.

D. Neighbourhood pruning

Although the two proposed definitions limit the neighbour space considerably, we can still find structures whose only difference is that one of them has more leafs hanging from some of the nodes. Thus, we prune every neighbourhood by leaving a maximum of two consecutive siblings of the same color and no children.

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.

Our dataset consists of all posts from March 2014 to May 2015 of the Podemos forum in Reddit[1][2]. 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)

A. Neighbourhood census

In this section we analyse the neighbourhood census of our dataset. A counter for every neighbourhood structure is created the first time we detect them. For every post, we extract its neighbourhood (order-based or time-based) and prune it as explained above (Section IV-D). Then check whether it is isomorphic to some of the already seen neighbourhoods. If this is the case, then we increment the counter for the neighbourhood previously seen. Otherwise we create a new entry for the current neighbourhood.

The census obtained with both type of neighbourhood are similar for the most frequent structures? That means that both kind of neighbourhoods are isomorphic for the most popular discussion structures.

B. Conversation-based clustering of users

In this section we categorize users based on the neighbourhoods of their posts. We limit our study to a set of 100 *active users* who wrote more than 100 posts. For those users, we create an initial feature matrix $U \times N$ where U is the number of users and N is the number of neighbourhoods in the census, and where the the position (u, n) is a counter of the number of times that a post written by user u has a neighbourhood isomorphic to n. We drop those feature columns that are zero for every user (these are neighbourhoods seen only around posts of users with low-level activity). To make the feature vector of a user independent on the number of posts, we transform the counts into percentages. And since some features have much higher percentages than others for most users, we scale and normalize the matrix so that every feature has mean 0 and variance 1. To avoid non-significant scores, we remove also the feature column corresponding to neighbourhoods that have a frequency less than 50 among the active users.

We use k-means to find the clusters, though one can use any other clustering method. Since the plot over the Within-Cluster Sum of Squares for $k = 1, ...20$ did not show any clear elbow we chose $k = 3$ clusters so that clusters are easier to interpret. We did the clustering over a feature matrix with order-based neighbourhoods and another feature matrix with time-based neighbourhoods. Figures **??** and shows the PCA projections of the users coloured by their assigned cluster.

VI. Conclusions

We presented a method to characterise conversations of users in online threads. Due to the tree nature of online threads, traditional patters such as triads are not able to capture much of relevant dynamics of a conversation. Our defined order-based and temporal-based neighbourhoods are able to capture a very rich variety of structures. We used this neighbourhoods to characterise users in terms of the structure of the conversations they participate in and showed that, indeed, there are different types of structural conversationalists

The concept of structural-temporal neighbourhood opens the door to some interesting paths of research. One might wonder whether other pruning are more pertinent than the proposed here. Also, even after pruning some neighbourhoods suggest the same type of conversation, so a manual merge might be convenient. Maybe sociologists can help to merge the found neighbourhoods in a meaningful way.

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

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

References

[1] R. Bhatt and K. Barman, "Global Dynamics of Online Group Conversations," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[2] R. Dorat, M. Latapy, B. Conein, and N. Auray, "Multi-level analysis of an interaction network between individuals in a mailing-list," *Annales des télécommunications*, vol. 62, no. 3-4, pp. 325–349, 2007. [Online]. Available: http://link.springer.com/article/10.1007/BF03253264

[3] S. Whittaker, L. Terveen, W. Hill, and L. Cherny, "The dynamics of mass interaction," in *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, 1998, pp. 257–264.

[4] N. Gaumont, T. Viard, R. Fournier-S'niehotta, Q. Wang, and M. Latapy, "Analysis of the temporal and structural features of threads in a mailing-list," in *Workshop on Complex Networks CompleNet*, 2016. [Online]. Available: http://arxiv.org/pdf/1512.05002v1.pdf

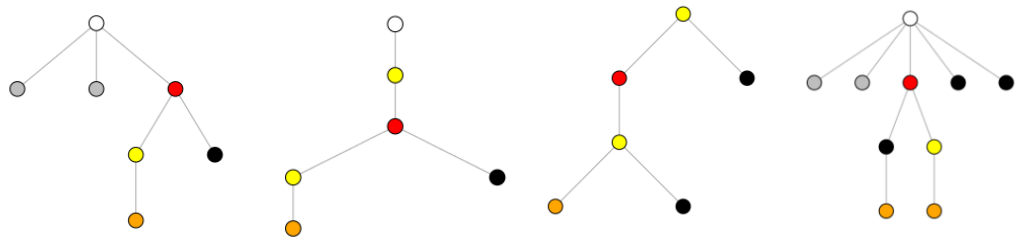


Fig. 3. Illustration of different real structures using different colors for the root (white), the ego (red) and other posts written by the same author (orange) the ego parent and other posts written by the same author (yellow), posts previous to the ego (grey) and posts posterior to the ego.

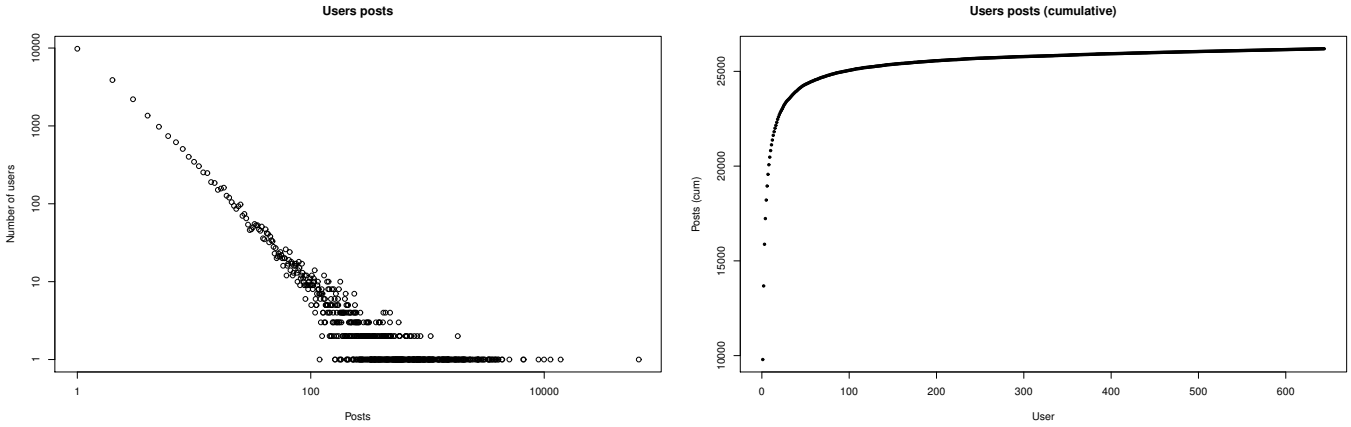


Fig. 4. Log-log and cumulative distributions and of posts per user.

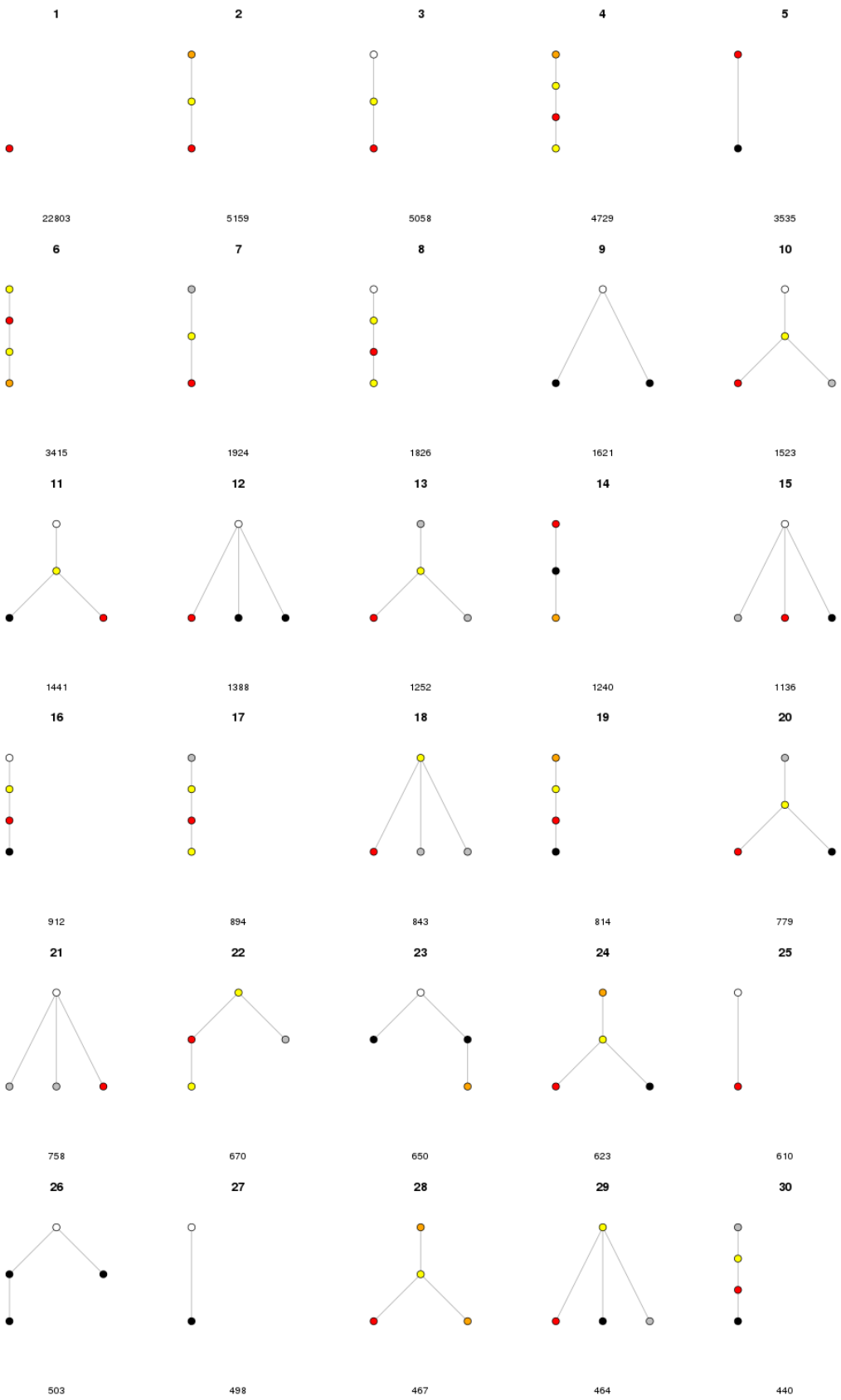


Fig. 5. Most frequent order-based neighbourhoods with $r = 2$ and $n = 4$

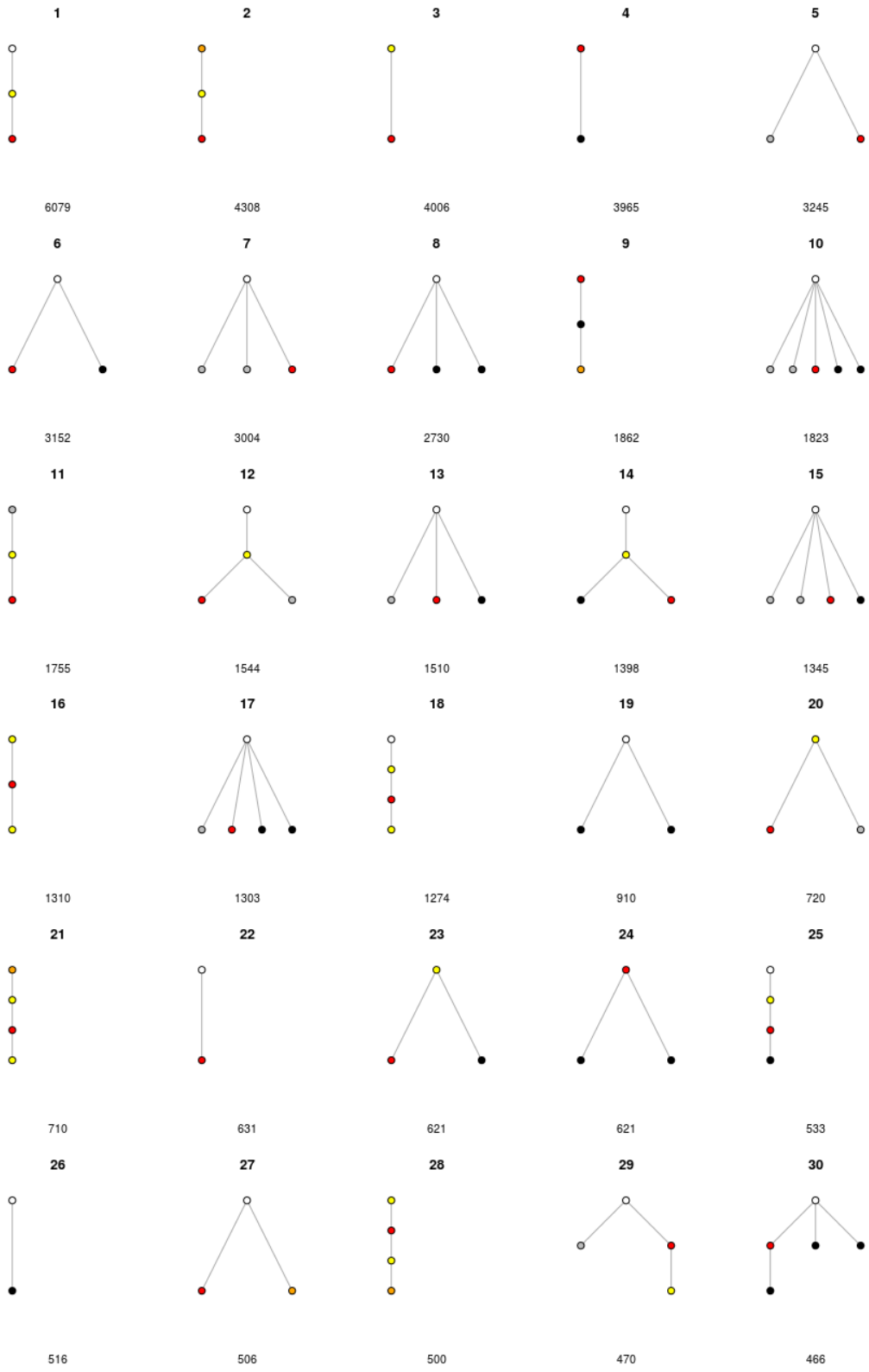
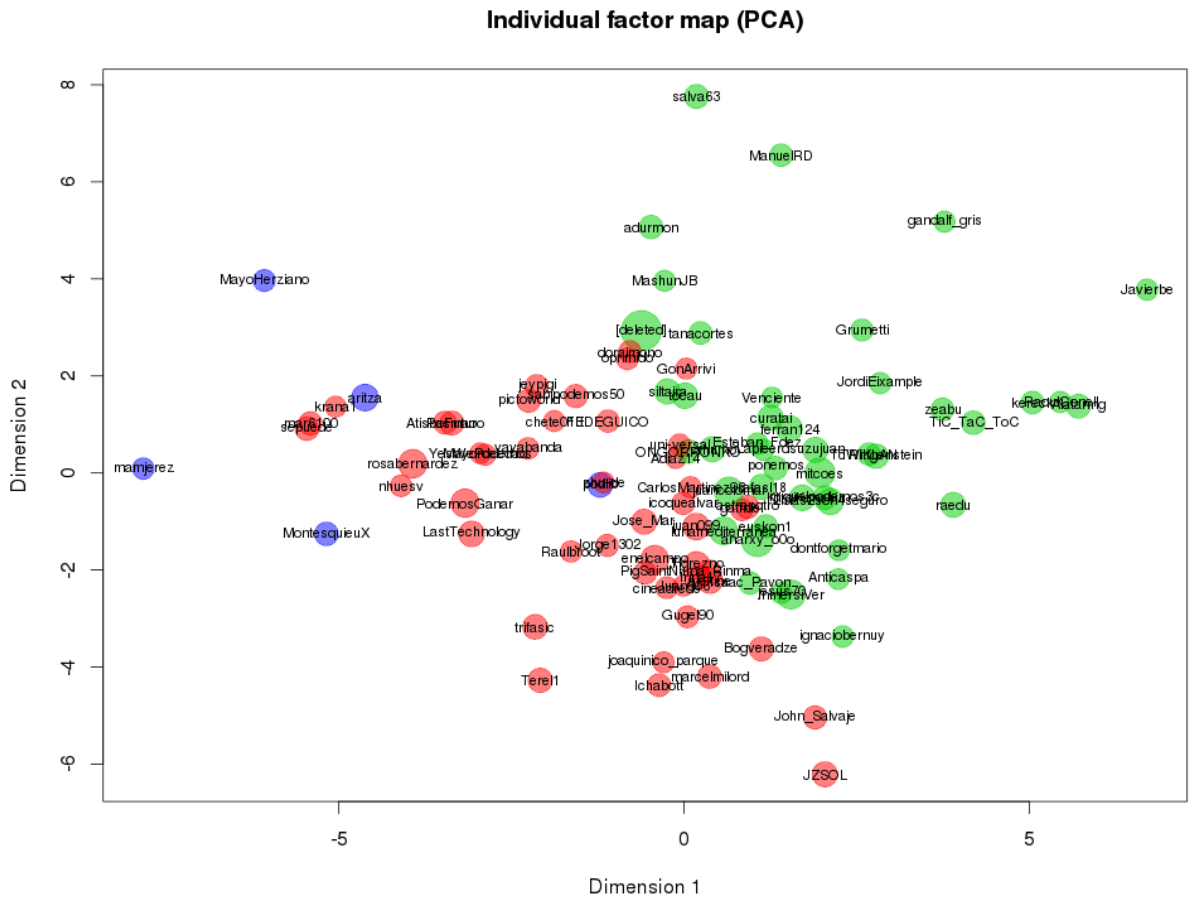


Fig. 6. Most frequent time-based neighbourhoods



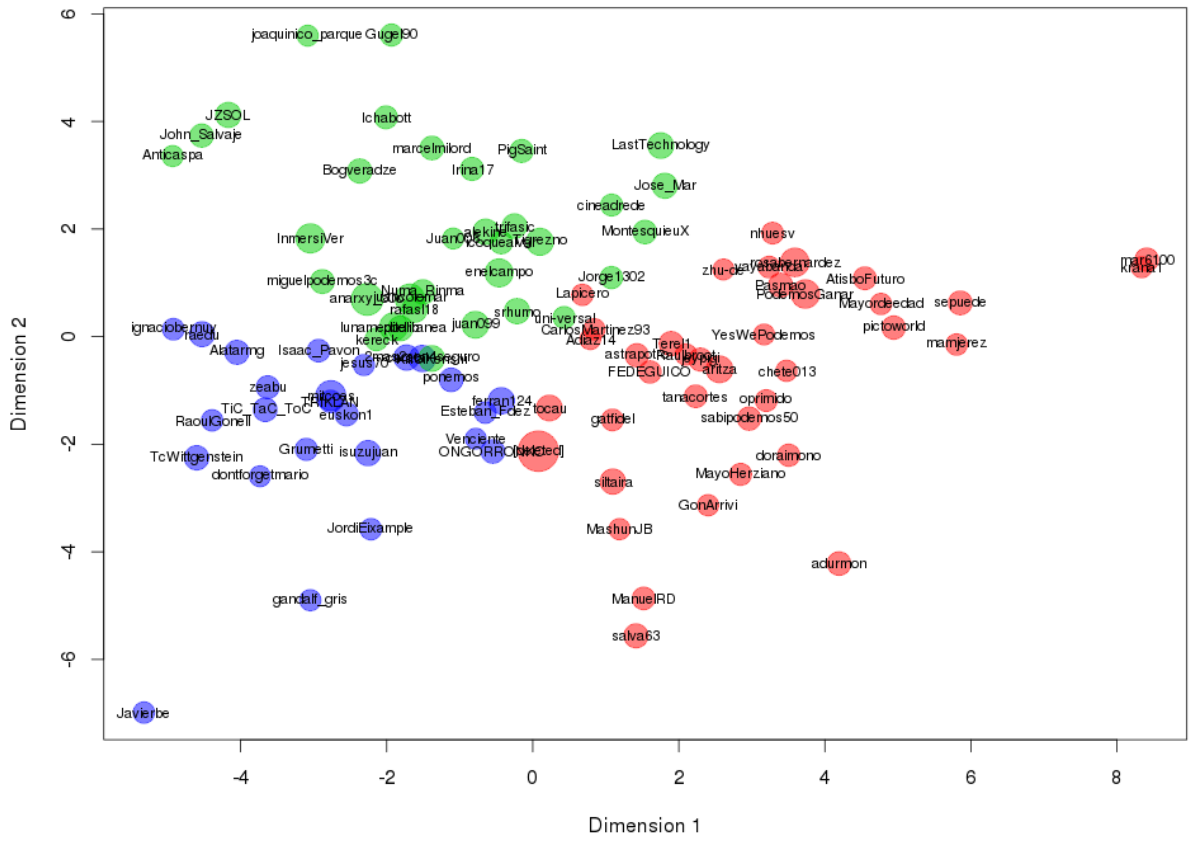


Fig. 7. PCA projections of the order-based (above) and time-based (below) neighbourhood features