Bloggers behavior and emergent communities in Blog space

M. Mitrović and B. Tadić

Department for Theoretical Physics, Jožef Stefan Institute, P.O. Box 3000, 1001 Ljubljana, Slovenia

Received 29 July 2009 / Received in final form 2 November 2009
Published online 17 December 2009 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2009

Abstract. Interactions between users in cyberspace may lead to phenomena different from those observed in common social networks. Here we analyse large data sets about users and Blogs which they write and comment, mapped onto a bipartite graph. In such enlarged Blog space we trace user activity over time, which results in robust temporal patterns of user-Blog behavior and the emergence of communities. With the spectral methods applied to the projection on weighted user network we detect clusters of users related to their common interests and habits. Our results suggest that different mechanisms may play the role in the case of very popular Blogs. Our analysis makes a suitable basis for theoretical modeling of the evolution of cyber communities and for practical study of the data, in particular for an efficient search of interesting Blog clusters and further retrieval of their contents by text analysis.

1 Introduction

In the information society, getting information from immediate users becomes an important way in opinion making, marketing products, and in everyday life issues as health, hazards, traffic problems, etc. In this respect the contemporary communication networks and social media offer on-line interactions, which facilitate communications, but on the other hand, they introduce a new kind of technology-mediated social clustering not known in the history. Behavior of users in the cyber space may be altered in comparison with common social interactions, which may induce new phenomena in the fast developing technology-based society. Obvious “absence of person” (the communication is indirect, mediated by text or video material posted on the web portals), accessibility, massive data available, and fast communications are some of the reasons which affect blogging at microscopic scale. The collective behaviors emerging in these interactions have not been yet understood. Therefore, scientific data analysis and modeling from the point of view of the complex evolving systems appears as a necessary step towards better understanding of the social phenomena in cyberspace.

Among variety of the currently available techno-social organizations, Blogs are somewhat specific, between the ordinary Web pages, where the owner writes its contents, and much faster information exchange between friends in FriendFeed, Facebook or MySpace networks. On the other hand, on Blogs the posts are short written forms, usually dedicated to a given subject and written by known author (or a group of authors) of the Blog. Other registered users can read and leave comments on the posts. Registration of users is required at many Blog sites and the action of each user is traced in time. The action of users is delayed compared with posting time. Apart from the subject, the act of writing the post and/or comment often requires certain dedication and writing skills, in which author’s personal profile and preferences get involved. In this way, posts may comprise an aesthetic as well as emotional and moral contents, similar to books, movies or music items accessible on the Web. Unlike friends networks, in-advance relationship between authors of Blogs and other users is not an underlying context on Blogs. This all might influence the evolution of Blogs, however, no precise measure of the importance of these and other factors exists so far. The mechanisms that drive user activity in the Blog space have not been fully understood [1].

Analysis of data in different social media revealed correlated behavior manifested in power-laws in the structure of networks and communities related to movie [2,3] and music genres [4,5], and cascades of events in the Internet forums [6] and through different Blogs via hyperlink connections between them [7–10]. Further regularities were found in the analysis of the Blog entries [11,12]. The approach with a targeted analysis of the contents of posted text on Blogs and in other forms [13,14], brings a new dimension in the study of the techno-social communications [15].

In this work we consider the data about actual events occurring between users-and-posts and map them to a bipartite network, in which users, as one partition, do not have a direct connection but interact through the posts, which makes the other partition. Links between users and posts are also directed, depending on the action (reading or writing the post/comment). In this way, our network
makes an enlarged Blog space, with users and posts (and comments to the posts) treated at the equal level. It is also related to a given Blog site (we consider two Blog sites with different organizational characteristics). In this enlarged Blog space we are able to study the post-mediated interactions between users and how the behavior of users affects the structure of the Blogs. Note that in several previous studies of Blogs [7,9,10] different networks have been considered: posts as the network nodes were connected with the hyperlinks pointing from one post to another, as the network edges.

Organization of the paper is as follows: in Section 2 we analyse the temporal features of both users and posts and determine statistical measures of users' activity relevant to the Blogs. In Section 3 we define the bipartite networks as emanating from the data and study their topological properties. In Section 4 we determine the community structure of users and fine clustering of posts within user groups. The difference between very popular and other posts is also pointed at the level of such communities. A short Summary of the results and conclusions are given in Section 5.

2 Temporal features of blogging

2.1 Data structure

We consider large Blog data, together about half of a million entries, collected from two different Blog sites, which have entirely different internal organization and history: BBC Blogs and Belgrade radio B92 Blogs. The case of B92 Blogs is interesting for the analysis for several reasons. First, we have collected all data from the beginning 27. May 2007 till 1. March 2009. Furthermore, on this Blog site users are registered not just to read and comment other posts, but to write their own post, and more importantly, no predefined categories of post subjects are imposed. Thus, the internal structure of posts emerges in a self-organized manner through user interactions on posts and comment-on-comment actions. The availability of posts is time limited to seven days (this rule was imposed after first few months of the functioning). Some of the users are upgraded to so called VIP authors, whose number fluctuates in time, and their recent posts are highlighted. Here we analyzed the posts written by all VIP users in the above mentioned period and consider all users related to these posts, which comprises of \( N_U = 4598 \) users, \( N_B = 4784 \) posts and \( N_c = 406527 \) comments to these posts. On the other side, the BBC Blogs exists for much longer time. For better comparison, we have collected the data form the same period as above, which gives \( N_B = 3792 \) posts, and \( N_c = 80873 \) comments written by \( N_U = 21462 \) registered users. In contrast to B92 Blogs, at the BBC Blogs both authors of posts and category of posts are predefined and fixed. The users are registered and allowed only to read and comment the posts. Accessibility of posts is not limited in time. In contrast to B92 Blogs, information about ID of a comment which is commented is not available in BBC Blogs, all comments are attributed to the original post.

2.2 User behavior in Blog space

Data about users, kept under their registered IDs, contain information about users’ activity over time, appearance of their posts, and their comments to other posts. In Figure 1 we show temporal patterns of the activity of more than 3000 users, ordered by the time of their registration on B92 Blogs within first year. As the figure shows, users appear in some waves, probably related to the external events or Blog site management, and newly registered users are active within some time intervals, whereas their activity is reduced in later times. Some users persist over long time period, while many other users either reduce frequency or stop writing on Blogs altogether. The heterogeneity in the users’ activity is further quantified by the analysis of time intervals between two successive user’s activities and by the number of events (written posts and/or comments) \( n_{com}(i,t) \) within a specified time window \( TW_{IN} \), for instance within one day. The distribution of time intervals between user’s successive activities is given in Figure 2 for both BBC and B92 Blog users. The power-law dependences over several decades (time is measured in minutes) suggest robust non-random patterns of users behavior, which is not much dependent on how the Blog site is organized. The slopes of the curves are different (1.5 and 1.15 at B92 and BBC Blogs, respectively), indicating slightly larger probability of large inactivity times at BBC Blogs.

Further quantitative analysis of the time series of activity, \( n_{com}(i,t) \), a given user \( i \) in time window \( t \), reveals additional information about user behavior. An example of such time series of a very active user from B92 is shown in Figure 3, with \( TW_{IN} \) equals one day. The power-spectrum of the time series shows long-range correlations at large frequency region, which suggests that the user activity is correlated over small time intervals.