Eur. Phys. J. B **77**, 597–609 (2010)
DOI: 10.1140/epjb/e2010-00279-x

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Networks and emotion-driven user communities at popular blogs

M. Mitrović[1], G. Paltoglou[2], and B. Tadić[1,a]

[1] Department of theoretical physics, Jožef Stefan Institute, Box 3000 SI-1001 Ljubljana, Slovenia
[2] Statistical Cybernetics Research Group, School of Computing and Information Technology University of Wolverhampton, UK

Received 6 May 2010/ Received in final form 25 August 2010
Published online 27 September 2010 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2010

Abstract. Online communications at web portals represents technology-mediated user interactions, leading to massive data and potentially new techno-social phenomena not seen in real social mixing. Apart from being dynamically driven, the user interactions via posts is indirect, suggesting the importance of the contents of the posted material. We present a systematic way to study Blog data by combined approaches of physics of complex networks and computer science methods of text analysis. We are mapping the Blog data onto a bipartite network where users and posts with comments are two natural partitions. With the machine learning methods we classify the texts of posts and comments for their emotional contents as positive or negative, or otherwise objective (neutral). Using the spectral methods of weighted bipartite graphs, we identify topological communities featuring the users clustered around certain popular posts, and underly the role of emotional contents in the emergence and evolution of these communities.[b]

1 Introduction

Science of the Web [1] is an emerging multidisciplinary area with interconnected contributions from the physics of complex dynamical systems, computer science, and social science. Apart from developing technology and algorithms for safe and efficient information processing, the research of Web concerns with understanding its structure [2] and the underlying evolution mechanisms [3] as well as the emergent social phenomena among Web users [4,5]. In this work we present a systematic methodology for study of the collective user behavior on Web portals. The approach is based on the physics of complex networks and the computer science methods of text analysis.

Emotions & Emerging Behavior in Cyberspace. Recent developments of the communication technologies have induced new types of human interactions mediated by the computer networks and on-line availability of different types of data. This makes the basis for new practice of social communications leading to potentially new technology-driven social phenomena not observed in conventional social mixing and thus calling for new science approaches [1,4,6,7]. On the other hand, huge amount of data of user communications over different Web portals is rapidly accumulating, which offers fabulous possibilities for the empirical study. The methodology of complex dynamical systems and mapping the data onto networks provides the ways to detailed quantitative analysis.

An important feature of the online communications is that *user interactions are mediated by the posted material*, e.g., the text of posts and comments on the Blogs, studied here. The indirect interactions not just change the conventional social rules known in face-to-face communication, but also indicates the importance of the contents of the posted material [8–10]. In the Blogs, the posted text may in different ways affect the behavior of the users who read it, depending on the information that the text contains, but also by featuring certain aesthetic, moral or emotional contents [8,11]. Recent studies increasingly show that the emotions expressed in the text (or other posted materials) play an important role in the online social dynamics. The strength of the emotions expressed by an individual, e.g., the user reading a posted text, can be measured in the laboratory [12] and observed on the level of large-scale social effects [11,13,14].

A number of conceptually different Web sites are currently available, ranging from the consumers opinion about products, e.g., movie database (IMDb), books and music records (Amazon), across the sites with exchange of opinions about everyday events (Diggs, Blogs, Forums), to fast on-line communication on friendship-based networks (Facebook, FriendFeed, MySpace). The Blogs are conceptually in between the consumer networks and the friends networks, mentioned above, and thus play a special role in the study of social on-line communities [8,9,15–19]. In Blogs authors express and exchange their opinion via written (short) texts, with other users, who are generally not acquaintances in real life. Registration of bloggers is

[a] e-mail: Bosiljka.Tadic@ijs.si

[b] All data are fully anonymized. No information about user IDs are given.

required on many Blogsites, which enables quantitative analysis and tracing user's activity over time.

Network representations of on-line interactions. Network representations of complex dynamical systems including social systems, has proved as a useful tool for quantitative study both in terms of the structure and the dynamics over networks (for a recent review see Ref. [20,21]). Mapping the data related to different social media onto networks reveals correlated dynamical behaviors, which is manifested in power-law dependences in the structure of networks and other related distributions [19,22–27]. The study of group formation in the networks related to movie data [22,28,29], music genre [24,25], subject of the posts in Blogs [19], forums [30], news sites and conference publication [31], etc., show that similar mechanisms might underline the behavior of humans in these on-line communications. Methods for analyzing content of short messages and textual posts [5,32] and their emotional content [12,33,34] enable understanding how the interactions on micro-level (user-to-post-to-user) leads to large-scale behavior within these virtual communities.

Mapping the data onto *bipartite networks* [19,22,25,28] is a suitable representation which enables the analysis and identification of different *user communities*. Statistical theory and community detection using the methods of the eigenvalue spectral analysis of networks reveal that different mechanisms may drive the dynamics on very popular post compared to all other posts. (Details of the spectral analysis of modular networks are described in Ref. [35], while other methods based on maximization of modularity are reviewed in Ref. [36,37]). In particular, the behavior of bloggers on normally popular posts [19] appears to follow a pattern of self-organized dynamical behavior and communities mostly related with the subject preference. Whereas, subjects appear completely mixed in the case of very popular Blogs [19], indicating different underlying mechanisms.

In this work we focus on studying *popular posts* collected from *bbc.co.uk/blogs/* by mapping the high-resolution data onto bipartite graphs and finding communities of users on it. We study the text of posts and comments of users within these communities with the aid of machine learning approaches, trained to detect and distinguish emotions in text. This enables us to study systematically the role of the emotions in the emergence and the evolution of the user communities and the patterns of user behavior at these popular posts.

2 Data structure and contents of popular blogs

We collected data [19] from the *bbc.co.uk/blogs/* site for time period of nearly two years, from June 2007 till February 2009. The dataset contains high temporal resolution of user IDs related action, posting comments related to a given post, as well as the IDs of the posts and comments and their text. The concept of the BBC Blogs is rather special: The original posts are written by few (invited) authors, who often do not take part in the discussion. All posts belong to one of the predefined categories, according to their subjects. Users are registered by IDs and allowed to make comments on these posts. The information about comment-on-comment is not stored, so that all comments are automatically attributed to the original post. The whole dataset consists of $N_P = 3792$ posts and $N_C = 80873$ comments written by $N_U = 21462$ users.

As mentioned above, we focus on the *popular* posts and analysis of the emotional contents of user comments related to them. As the popularity break-point occurs at the number of comments $\gtrsim$100 (see the discussion below and Ref. [19]), from the entire dataset we select these posts and all users and their comments related to them. We find $N_P = 248$ popular posts and $N_U = 13\,674$ users who wrote $N_C = 53\,606$ comments on these posts. We downloaded text of each of these posts and text of each related comment, and analyzed it with the *emotion classifier*, described below. These posts appear to belong to five different subject categories: *Business and Economy*, *Music and Art*, *Sport*, *Technology* and *Nature and Science*. Knowing the authors and the posting times for all posts and comments, we are able to reconstruct temporal patterns of users behavior and link it to the emotional contents of the texts.

2.1 Mapping the data onto bipartite networks

The Blog data can be suitably represented by directed bipartite graphs with *users* as one partition, and *posts and comments*, as the other partition [19]. By definition [38], in bipartite networks links are allowed only between nodes of different partitions, which completely respects the structure of the interaction between users over posts and comments in the Blog data. In the data we have $i_U = 1, \ldots, N_U$ users and $j_B = 1, \ldots, N_P + N_C$ posts and comments, which together make $N = 106\,127$ nodes of the bipartite network which eventually is reduced to $N = 67\,528$ nodes in the case of the popular posts. The post/comment j_B is linked to its author i_U trough a directed link that points form user to posts/comment, ($i_U \rightarrow j_B$). A directed link in the opposite direction ($i_U \leftarrow j_B$) indicates that the user i_U left a comment on the post j_P. (Note that one user can write more than one comment to a specific post, resulting in the multiple outgoing links). Following these rules we obtain a *directed bipartite network representation of the blog data*. An example of a single-post network, which illustrated these rules is shown in Figure 1a. Together with the information about time of the appearance of each user, post, comment and the link, the network contains full information from the dataset. Applying the graph theory methods we can now study the structure of the interactions at different levels, from individual nodes of both partitions, to the mesoscopic (community) structure to the level of the entire network, as well as the evolution of the network. We utilize the emotional content of each post and comment as an additional feature that interferes with the network structure. It is

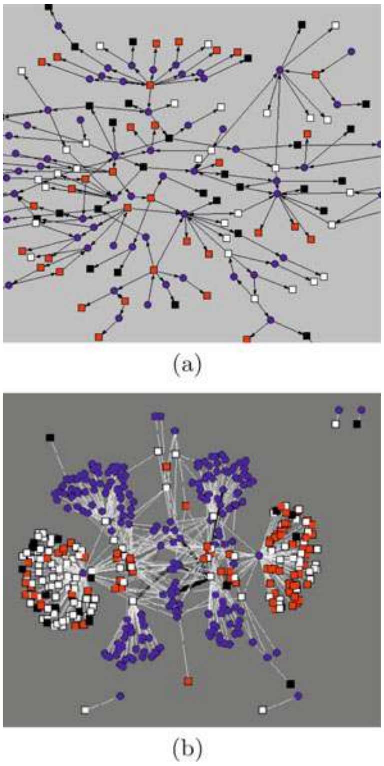

Fig. 1. (Color online) Example of a directed bipartite network with users (circles) and comments (squares) related to a particular post (a) and a weighted symmetrical posts-and-users network (b). Color of the posts and comments indicates their emotional content classified as positive (red), negative (black) or neutral (white).

indicated by the color of the comments (see detailed discussion below). Note that these networks can become very large, depending on the dataset considered.

Other suitable representations can be obtained by *compressing* or *projecting* from these bipartite graphs. In reference [19] we have studied several networks of Blog data obtained by projecting of the bipartite networks onto *user-projection* and *post-and-comments projection*. For the purpose of this work, i.e., the networks of popular posts, we consider a compression of the whole network into a *weighted bipartite network*, which consists of *users* and *posts* only, while the weights of the links between them is given by the number of comments that the user left on the related post. An example of such network is also given in Figure 1b, together with the color, that indicates the cumulative emotional content of all comments related to the post.

2.2 Extracting emotions from text sentences

We view the problem of extracting the emotions from text sentences as a *classification problem*. The general aim of classification is, given a document[1] D and a fixed set of classes $C = \{c_1, c_2, ...c_t\}$, to assign D to one, or more, of the available classes.

We have implemented two supervised, machine-learning classifiers for estimating the probabilities whether a document D is objective or subjective, positive or negative. The classifiers function in a two-tier fashion: the first-stage classification determines the probabilities of whether D is objective of subjective, i.e., $C_1 = \{obj, sub\}$, and the second determines the probabilities of the polarity of the document, i.e., $C_2 = \{neg, pos\}$, if it was classified as subjective in the first-tier classification. The final output of the classifiers is therefore one of $\{obj, neg, pos\}$.

We have utilized language model classifiers [39,40] for both classification tasks. The aim of the classifiers is to maximize the posterior probability $P(c|D)$, that a given document D belongs to class c. Typically, the best class is the *maximum* a posteriori (MAP) class c_{MAP}:

$$c_{MAP} = \arg\max_{c \in C} \{P(c|D)\}. \tag{1}$$

Using Bayes rule, we get:

$$\begin{aligned} c_{MAP} &= \arg\max_{c \in C} \left\{ \frac{P(D|c) * P(c)}{P(D)} \right\} \\ &\propto \arg\max_{c \in C} \{P(D|c) * P(c)\}. \end{aligned} \tag{2}$$

Furthermore, we have removed the denominator $P(D)$ since it does not influence the outcome of the classification. $P(c)$ is the prior that indicates the relative frequency of class c, i.e., all other things being equal, the classifier will prefer the most frequent class.

Language models operate by estimating the probability of observing document D, given class c. We represent D as token sequence $\{w_1, w_2, \ldots w_n\}$, therefore the aim of a language model is to estimate the probability of observing the above sequence, given c:

$$\begin{aligned} P(D|c) &= P(w_1, w_2, ...w_n|c) \\ &= P(w_1|c) * P(w_2|c, w_1) * ... * P(w_n|c, w_1, w_2, ..., w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|c, w_1, ..., w_{i-1}). \end{aligned} \tag{3}$$

Usually, an *n-gram* approximation is used to estimate equation (3), which assumes that the probability of token w_i appearing in document D depends only on the preceding $n-1$ tokens:

$$P(w_i|c, w_1, \ldots, w_{i-1}) = P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}). \tag{4}$$

A straightforward way to calculate the maximum likelihood estimate of $P(w_i|c, w_{i-(n-1)}, ..., w_{i-1})$ during the training phase of the classifier, given a set of documents and their respective categories, is by counting the frequency of occurrences of the tokens sequences:

$$P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{\#(c, w_{i-(n-1)} \ldots w_i)}{\#(c, w_{i-(n-1)} \ldots w_{i-1})}, \tag{5}$$

[1] The term "document" is used here in the broadest of senses, signifying a sequence of words. In realistic environments, "document" can be any sort of textual communication between two or more parties, such as blog posts, forum comments or Instance Messaging utterances.

where $\#(c, w_{i-(n-1)} \ldots w_{i-1})$ is the number of occurrences of token sequence $w_{i-(n-1)} \ldots w_{i-1}$ in documents of class c during the training phase and $\#(c, w_{i-(n-1)} \ldots w_i)$ is respectively the number of occurrences of sequence $w_{i-(n-1)} \ldots w_i$ in documents of class c during the training phase.

Despite the simplification that we introduced for the estimation of $P(D|c)$ using *n-grams*, the probabilities of actually observing a significant number of times, any specific sequence of n tokens in the training set is susceptible to the *sparse data* problem, especially for large values of n (i.e. usually more than 3) and small to medium-sized training corpora. Simply put, high order *n-grams* do not occur often enough to provide a strong indication of class preference.

For that reason, we usually further break the estimation of probability $P(D|c)$ to smaller *n-grams*, in a process that is called *smoothing*. There are various methodologies for smoothing [40], such as Laplace, Good-Turing or back-off estimators, etc, but in this work we adopted the Witten-Bell approach [41]. Therefore:

$$P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}) = \lambda_{(c,w_{i-(n-1)},\ldots,\, w_{i-1})} * P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}) + \left(1 - \lambda_{(c,w_{i-(n-1)},\ldots,\, w_{i-1})}\right) * P(w_i|c, w_{i-(n-2)}, \ldots, w_{i-1}), \tag{6}$$

where

$$\lambda_{(c,w_{i-(n-1)},\ldots,w_{i-1})} = \frac{\#(c, w_{i-(n-1)}, \ldots, w_{i-1})}{\#(c, w_{i-(n-1)}, \ldots, w_{i-1}) + L * W_{(c,w_{i-(n-1)},\ldots,\, w_{i-1})}}. \tag{7}$$

Here $W_{(c,w_{i-(n-1)},\ldots,\, w_{i-1})}$ is the number of extensions of the specific token sequence:

$$W_{(c,w_{i-(n-1)},\ldots,\, w_{i-1})} = \left|\left\{w_k|\#(c, w_{i-(n-1)}, \ldots, w_{i-1}, w_k) > 0\right\}\right| \tag{8}$$

and L is the *hyperparameter* of the distribution. Its aim is to provide a balance between higher and lower order n-grams. A high value of L gives more weight to lower n-grams, which is usually appropriate for smaller training sets, where the probability of encountering a higher order n-gram is small. In the reported experiments, the value of L was set to the value of the longest n-gram.

The algorithms for training and applying the classifier are presented in pseudocode below.

We experimented with unigrams and bigrams ($n = 1, 2$, respectively) which provide an acceptable compromise between effectiveness and efficiency. We trained the language model classifiers on the BLOGS06 dataset [42,43]. The dataset is comprised of an uncompressed 148GB crawl of approximately 100 000 blogs and their respective RSS feeds. The dataset has been used for 3 consecutive years by the Text REtrieval Conferences (TREC)[2].

[2] www.trec.nist.gov

Algorithm 1 Train Language Model classifier

INPUT: Documents $D = \{d_1, \ldots, d_n\}$, Classes $(C = \{c_1, \ldots, c_t\})$, Function $\Phi : D \times C \to \{T, F\}$, Value of n
$W \Leftarrow \{w_{k-(n-1)}, \ldots, w_k | d_i = \{w_1, \ldots, w_{k-(n-1)}, \ldots, w_k, \ldots, w_n\} \in D\}$
for all $c \in C$ **do**
    $|D_c| \Leftarrow |\{d_i|d_i \times c \to \{T\}\}|$
    $P(c) = |D_c|/|D|$
    $W_c \Leftarrow \{w_{k-(n-1)}, \ldots, w_k | \Phi(d_i, c) = T\}$
    for all $w \in W_c$ **do**
        Count $\#(c, w)$
    end for
    for all $w \in W_c$ **do**
        $P(w|c) \Leftarrow$ Equation 6
    end for
end for
return $n, W, P(c), P(w|c)$

Algorithm 2 Apply Language Model classifier

INPUT: $n, W, P(c), P(w|c)$ from Algorithm 1
INPUT: Document $d = \{w_1, w_2, \ldots w_n\}$ to be classified
$W_d \Leftarrow \{w_{k-(n-1)}, \ldots, w_k | d = \{w_1, \ldots, w_{k-(n-1)}, \ldots, w_k, \ldots, w_n\}\}$
for all $c \in C$ **do**
    $Score(d) \Leftarrow P(c)$
    for all $w \in W_d$ **do**
        $Score(d) += P(w|c)$
    end for
end for
return $arg\max_{c \in C} Score(d)$

Participants of the conference are provided with the task of finding documents (i.e. blog posts) expressing an opinion about specific entities X, which may be people, companies, films etc. The results are given to human assessors who then judge the content of the posts and assign each one a score: "1" if the document contains relevant, factual information about the entity but no expression of opinion, "2" if the document contains an explicit negative opinion towards the entity and "4" is the document contains an explicit positive opinion towards the entity. We used the produced assessments from all 3 years of the conference to train our classifiers, resulting in 150 different entity searches and 16 481 documents with a score of "1", 7 930 documents with a score of "2" and 9 968 with a score of "4", which were used as the "gold standard" for training our classifiers.

Specifically, for the first-level classification (i.e. $C_1 = \{obj, sub\}$) we used the documents that were given a label of "1" as objective and the union of "2" and "4" as subjective. For the second stage classifier (i.e. $C_2 = \{pos, neg\}$), we used the documents assigned a label of "2" as negative and "4" as positive. The resulting classifiers have an accuracy of approximately 70% for either classification task, using 10-fold cross validation.

3 Emergent structure of networks at popular blogs

3.1 Topology of bipartite networks and their projections

Topological properties in bipartite networks, e.g, the *degree* of a node and *commons* defined for pairs of nodes, as well as *community* structure [19] and the *mixing* patterns [28], are quite different compared with their monopartite projections and depending on the partition to which the considered nodes belong to. In particular, the *degree distribution* representing the number of links per node in the case of our bipartite networks has a special meaning in each partition. For the user partition, the node degree represents the number of comments left by that user. While in the posts partition the degree of a node is determined by the number of users who left comments on it. Also incoming and outgoing links are distinguished. The results are shown in Figure 2 for the entire dataset described in Section 2. These degree distributions exhibit a broad power-law functional dependences [19]. The distributions of *commons*, representing the common number of users per pair of posts C^U_{ij}, and vice versa, the common number of posts per pair of users C^B_{ij}, appear to have power-law dependences in different Blogs datasets, see Figure 2. Note that the corresponding commons C^B_{ij} or C^U_{ij}, appear as the weighted links in the monopartite projections on user or post and comments network, respectively. The power-law distribution of commons thus indicates strong inhomogeneity of the weights both in the user and in the post networks, indicating different importance of the nodes. Thus, it can be used in reducing the size of these networks, for instance according to the *strength* of nodes.

In reference [19] we have analyzed in detail the topological properties of the bipartite networks obtained by the mapping of the data of Blogs. Note that due to the absence of information about comment-on-comment in the BBC Blogs, the outgoing links are attributed only to the posts. Every post or comment can have only one author, resulting in only one incoming link per post/comment. Different structure is found in the B92 Blogs in which comment-on-comment is preserved in the data [19]. The out-degree distribution function for the BBC posts, which is relevant for the discussion in this work, is shown in Figure 2a. The distribution exhibits two slopes, indicating the existence of two separate group of posts: less popular posts related to a curve with smaller slope, and popular posts related with the tail of the curve. The breakpoint corresponds to $n^*_{com} \sim 100$ comments per post. Similar features are found in the case of B92 Blog [19] and movies [28], where behavior of users related to popular posts (movies) is qualitatively different compared to less popular posts, both in terms of the subjects and their mixing patterns.

As stated above, for further analysis in this work we consider only the popular posts with $n_{com} > 100$. From the whole corpus of data we first separate the data related to the popular posts, which includes the user IDs and times of their actions as well as the full text of all their comments left on these posts. The selected data will be analyzed below in terms of the networks and emotional contents of user comments. We first determine the topological community structure on these networks and then analyse the emotional contents of the comments within the community, using the emotion classifier described above in Section 2.2.

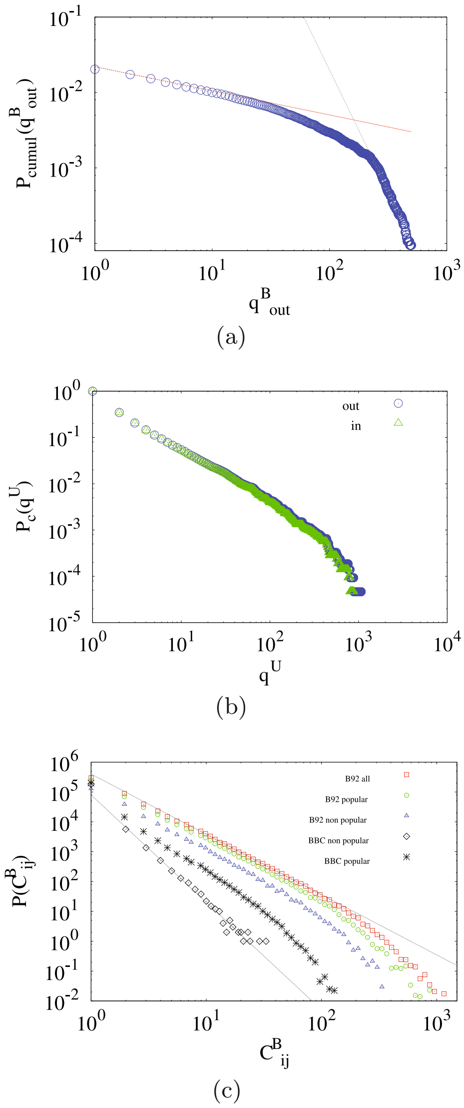

Fig. 2. (Color online) Out-degree distribution of all posts from BBC Blog dataset, indicating a breakpoint for the popular posts (a) and cumulative distribution of user-degree on these posts (b). Distributions of common number of posts per pair of users (c). ((c) reprinted from Ref. [19], ©Springer 2009).

3.2 Identifying user communities by spectral analysis of networks

In analogy to the social groupings in real life, identifying the emerging communities among users in Cyberspace represent a key point in the analysis of collective techno-social phenomena. The bipartite network representation of Blog data, as explained above, is a good basis for the quantitative analysis of user communities. In previous work [19] we have shown how the communities can be identified using the topology of these bipartite graphs. Specifically, the bipartite graph is first *projected onto user partition*, in which way we obtain a weighted and highly clustered networks of users. The weight of the links in these user networks is given by the above mentioned *commons*, C_{ij}^B, the number of common posts per each pair of users (ij). Then the communities, i.e., subgraphs on these networks that can be identified by topological and/or weighted structure of links, are searched within these user network. Note that most of the standard methods for the community structure analysis (see recent reviews in Refs. [37,44]) have a weak point when the weighted or strongly clustered networks are considered. As discussed above, the user networks on Blogs are both weighted and strongly clustered. For these reasons we use the eigenvalue spectral analysis methods, that works well in the case of clustered networks [19,35], however, it can not be applied to very large networks. See also the methods based on the idea of the maximum likelihood, which is adapted for the weighted graphs in reference [45]. Other methods suitable for weighted directed graphs, based on information theory and random walks dynamics, are discussed in reference [46].

In the present case, the bipartite network of users and popular posts, consisting of $N_U + N_P = 13674 + 248$ nodes is projected onto user partition by the procedure described above (see also [19,28]). Since the weight of the link between two users is given by the number of their common posts has power-law distribution, shown in Figure 2, the network can be considerably reduced by cutting the links with small weight. Specifically, keeping the links with weight $C_{ij}^B > 1$, i.e., keeping the *users who have commented more than one post within a community*, we reduce the network to $N_U = 3592$ users. We then apply the eigenvalue spectral analysis of the reduced network and identify its communities.

Finding communities in networks is based on certain properties of the eigenvalues and eigenvectors of the adjacency matrix and other matrices related to a network structure which has topological subgraphs. We use normalized Laplacian matrix related to the diffusion dynamics on networks [35,47], defined as

$$L_{ij}^U = \delta_{ij} - \frac{C_{ij}^B}{\sqrt{l_i l_j}}, \tag{9}$$

where C_{ij}^B is the weighted link of the user projected network, as explained above. The normalization factor in equation (9) l_i is the *strength* of user i, which is defined as sum of weights of all links at this node $l_i \equiv \sum_j C_{ij}^B$. In the networks emerging from the Blogs data the strengths also obey power-law distributions.

Detailed description of the spectral method can be found reference [35], here we briefly highlight the relevant properties of normalized Laplacian, given in equation (9).

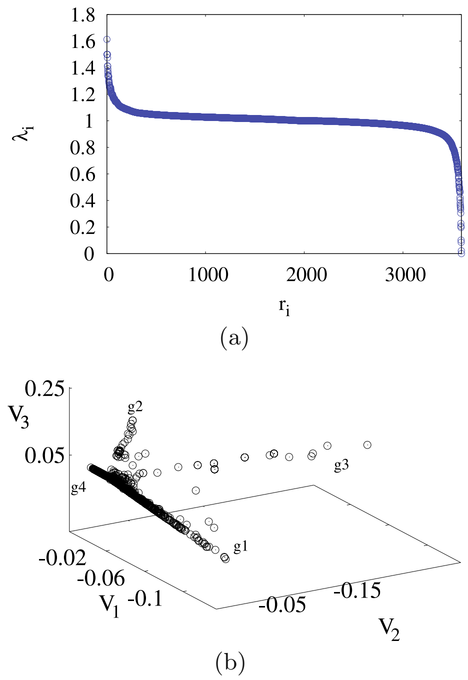

Fig. 3. (Color online) For the weighted user-projection of the bipartite network of the popular BBC posts: spectrum (a) and 3-dimensional scatter plot of the eigenvectors corresponding to three smallest non-zero eigenvalues (b), indicating four distinct user communities.

Since the weighted projection of the bipartite network on user partition is undirected, the normalized Laplacian has a symmetric form, thus resulting in real eigenvalues and orthonormal set of eigenvectors. The spectrum of the normalized Laplacian is limited in the range $[0, 2]$ with one zero eigenvalue in the case of a connected network. The set of the lowest non-zero eigenvalues is separated from the rest of the spectrum and the number of these eigenvalues coincides with the number of topologically distinct subgraphs, i.e., communities in the network [19,35,48]. Moreover, the eigenvectors corresponding to these eigenvalues have a nonzero positive/negative components localized on the subgraphs. This property is compatible with a branched structure when these eigenvectors are plotted against each other (see Fig. 3 for the present case). In this way, in a suitably selected projection, one can identify a community as a separate branch in the scatter plot of the eigenvectors. IDs of users belonging to each of the branches, e.g., in Figure 3, are then collected and the list of their posts and comments are selected from the dataset.

The spectral analysis of the weighted network of BBC bloggers related to the posts of normal popularity (i.e., number of comments below the breakpoint in Fig. 2a) is given in reference [19]. Closer inspection of the posts related to the users in each community shows that they are clustered according to the posts subjects. However, the communities identified here in Figure 3b on popular posts, do not entirely follow this pattern. Among four groups shown in Figure 3b, we find that the largest group g_4 consists of users who often comment different types of posts,

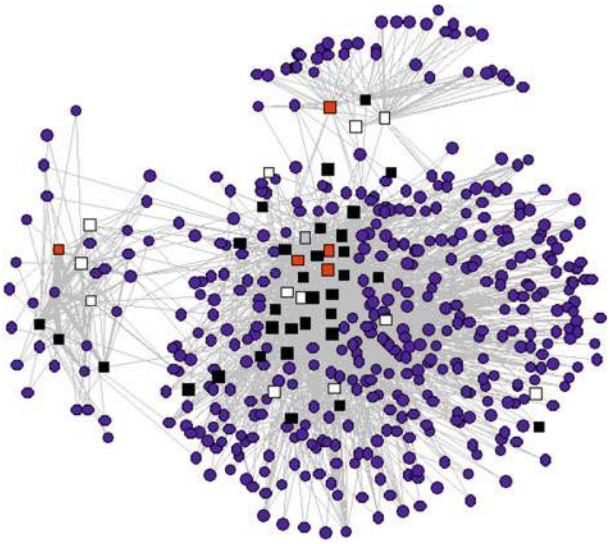

Fig. 4. (Color online) An example of the weighted bipartite network of users (circles) and posts (squares) from the BBC popular Blogs. The weighted links represent the number of comments left by the user on the related post. The color of the posts signify the average emotional content on the post, computed from all comments left on that post: negative (black), neutral (white), and positive (red). Note that few posts, marked by gray color, contain video material and were not classified.

while the remaining three groups (in the plain orthogonal to g_4 branch) are formed by users with interest in specific posts: g_1, g_2 and g_3 consists of users that commented mostly posts on different Sports. Behavior of the users related to popular posts is expected to be different and driven by emotions [19]. Existence of group of very active users that leave a comment on most of the posts from popular group is also found in B92 Blog networks [19] and in the network of movie users [28], suggesting another universal pattern of user behavior at popular posts.

In Figure 4 three of the four user groups identified above are shown together with the posts which they are commenting. This is an example of the compressed bipartite network with the number of comments of a user to a post is given by the width of the link between them. By color on the post nodes is shown also the average emotion of the post. It is computed from the emotional contents of all the comments left at that post, and averaged over the number of comments. The averaging will be explained below in Section 4.1, where the emotional contents of individual comments is precisely defined and determined. We also give more details of the dynamics of the emotions in the next section.

4 Evolution of communities and emotional contents of the popular blogs

4.1 Emotional contents of popular posts

We performed text analysis of all 242 popular posts and all related comments using the language model classifier, as described above in 2.2. The probability that a given text x is objective, $P_{obj/x}$ computed from its text is compared with a threshold of objectivity $C_{obj/x}$, which for the present analysis we set to $C_{obj/x} = 0.43$. That is, a particular text is classified as *objective*, i.e., having value of emotional content equal 0, if $P_{obj/x} > C_{obj/x}$. Notice that this means a large threshold value for the criteria of *subjectivity* of texts.In this way selected *subjective* texts are further classified against *positive* and *negative* emotional content by the second stage of the classifier, as explained above, and the probability $P_{pos/x}$ is determined for each particular text. Since the classifier seems to be biased towards positive comments, the threshold for positive emotions is adjusted so that the emotional content of the post/comment is +1 if $P_{pos/x} > C_{pos/x}$ and −1 otherwise, with $C_{pos/x} = 0.7$. The bias is a result of the unbalanced data set that was used for the training of the classifiers, i.e., 7930 negative vs. 9968 positive documents. The thresholds were chosen after an exhaustive search of the parameter space in order to maximize the accuracy of the classifiers for each classification step.

Having determined the emotional content of all textual posts and comments in our popular posts dataset, we perform further analysis in order to infer the role of emotions in the blogging dynamics. The distributions of the number of comments per post that are classified as objective, negative or positive, in our dataset are given in Figure 5a. It shows that the tail in the distribution develops, which is more pronounced in the case of negative comments. In the dataset, about 50% of all comments and posts appear to have negative emotions, while less than 25% of them are classified as positive.

Based on the emotional content of the comments related to each specific post, we define two variables, *emotional charge* Q^B and *total number of emotional comments* Q^B_ν, whereby we can measure and describe the emotional state of a particular post in different time intervals. In particular,

$$Q^B_\nu(t) = N_{c+}(t) + N_{c-}(t); \quad Q^B(t) = N_{c+}(t) - N_{c-}(t); \tag{10}$$

where $N_{c\pm}(t)$ stand for the number of positive/negative comments on a given post at time t after its posting. Temporal evolution of these quantities over lifetime of a post differ for different popular posts. Specifically, we find two typical patterns with (a) the number of (emotional) comments bursting soon after the posting, and (b) the number of emotional comments steadily increases over time. In the first case the pattern also shows a burst of the negative charge (excess of the number of negative comments). Whereas, a weakly negative charge is detected in the case where the number of comments increase slowly. The situation is illustrated in Figure 5b with two examples of posts, representing these different evolution patterns. Specifically, these are the posts named "College 5" ($MA1227829$) from the category *Music and Arts* and "Woolworth into administration" ($EB144093$) from *Business and Economy*. Closer inspection into evolution and emotional contents of the comments on these two posts reveals that the emotional comments on the economy post

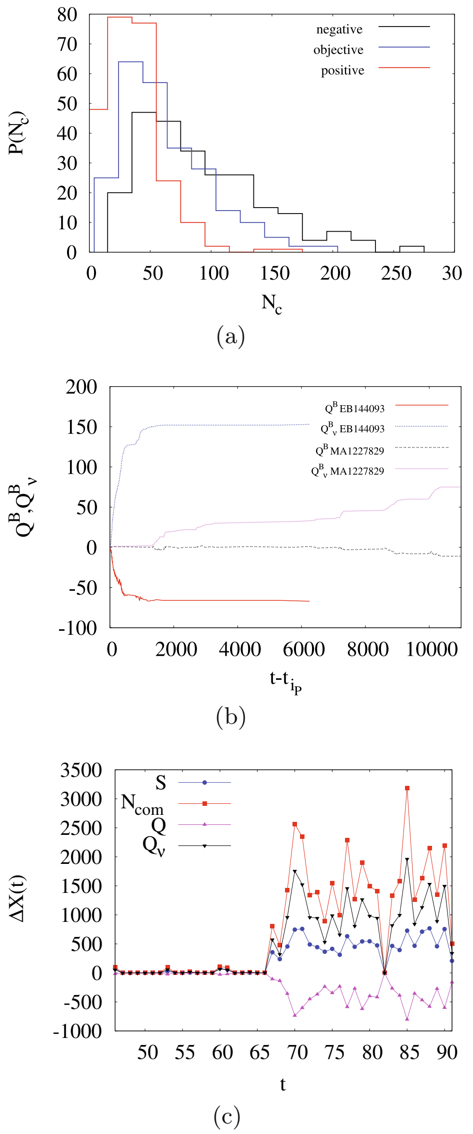

Fig. 5. (Color online) (a) Distribution of positive, objective and negative comments on one post averaged over all 242 textual posts in the dataset. (b) Evolution of the number of emotional comments Q_ν^B and charge Q^B plotted against time since posting, in two selected post with different patterns of popularity. (c) For the large community from Figure 4: fluctuations of the size of community plotted against time in weeks. Shown are also the fluctuations of number of all comments, number of emotional comments and charge of the posts.

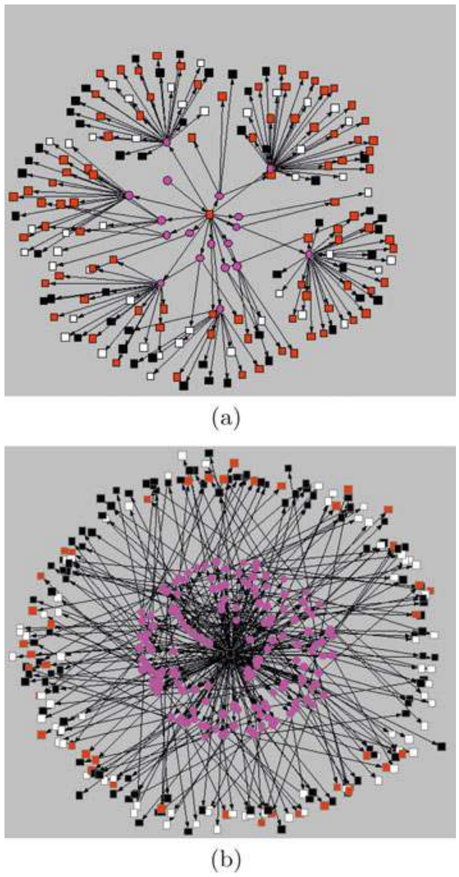

Fig. 6. (Color online) Two types of "popularity": network of a discussion-driven popularity, music post "College 5" (a), and of an externally-triggered popularity, the post is from the economy and business category: "Woolworth into administration" (b).

and the observed excess negative charge (critique) sharply increases over first two days and then stabilizes. While, in the case of music post, the number of emotional comments steadily increasing with the overall emotional content balancing around zero. The networks of users and comments at these two posts are shown in Figure 6, clearly indicating *two types of popularity*: on one side, few users exchanging many comments over time, as in the case of the music post, while on the other, many users are posting one or few comments, as in the case of economy post. The comments classified as positive/negative or neutral are indicated by the color, red/black or white. Apart from different evolution and the emotional content, these two patterns of popularity can be related with different mechanisms with discussion driven, and externally driven evolution, respectively.

The communities found above are formed of users who are linked to one or more such popular posts. An example illustrating connection between the network structure and expressed emotions in the text of posts is the weighted bipartite network shown in Figure 4. It contains users belonging to three groups, g_1, g_2 and g_3, found by spectral method (cf. Fig. 3) and the popular posts to which they are linked, altogether $N_U = 424$ users and $N_P = 50$ posts. A link between post i_p and user j_U on this network means that user has left a comment on that post. Multiple linking (weighted link) occurs if the user left more than one comment on the same post. For every post we calculate the average emotional content according to $E_{i_P} = \frac{1}{Q_\nu}\sum_{i_c \in S_{i_P}} e_{i_c}$, where e_{i_c} is the emotional content of the comment i_c ($e_{i_c} \in \{-1, 0, 1\}$) and S_{i_P} refers to the set of emotional comments left on post i_P, its size is equal to parameter Q_ν^B of that post. Therefore, the color of a post represents it average emotional content

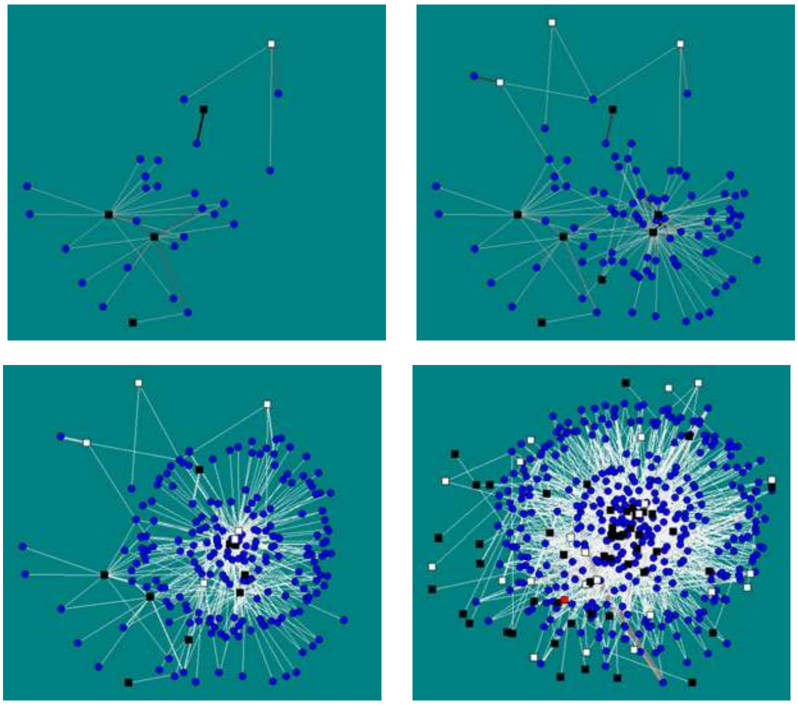

Fig. 7. (Color online) Snapshots of four stages in the evolution of a community of users (circles) related to several popular posts (squares) at *bbc.co.uk/blogs/*. The overall emotional charge of the posts is indicated by the color.

after the whole evolution time (within the dataset). A threshold for neutrality is applied for the average value $-0.25 < E_{i_P} \leq 0$, as explained above in relation to Figure 4 . (Note that few gray posts occurring in Figure 4 are the posts with video content, which can not be classified by our emotion classifier). The color of user nodes is not related to any specific property.

As one can see the network in Figure 4 has community structure, which is expected based on the above spectral analysis. Every group of users is grouped around a core of the few posts. Some posts are commented by the users from different groups and have special connecting role in the network. Figure 4 shows that the largest community of users is grouped around the core of mostly negative posts, only few posts in this group appear to have in the average positive or balanced emotional content. Note that in this representation as a weighted bipartite network, these posts actually keep the whole community of users together. By removing the post nodes, the community would literally desolve into individual user nodes!

The community structure on Blogs is not fixed but evolves over time. We looked in detail how the largest community in Figure 4 evolved, staring from the appearance of the first post. The *fluctuation of size* of the community, i.e., the number of active users within a time bin of one week, is shown in Figure 5c, for the whole span of time in the dataset. Shown are also the time series corresponding to the fluctuations in the number of comments that the active users left, the fluctuations in the number of emotional comments and fluctuations in the charge of these comments over the same time periods. It is remarkable that the increase in the size of the community is closely related to excess of negative comments, and vice versa, the community decreases when the negative charge vanishes.

An example of the evolution of a user community related to popular BBC posts is visualized as an evolving weighted bipartite network in Figure 7. It indicates that only certain among several posts become very attractive and get commented by many users. It also shows that the average emotional contents of the related popular posts are predominantly negative. Emotional contents of the posts and comments have impact onto further user's action. Within our datasets we are able to follow user conduct over time. In the following section we study the patterns of user behavior, in particular in connection with the emotional contents of the texts.

4.2 Patterns of user emotional behavior at blogs

Full information about user's action over time and the contents of its comments give an opportunity to analyze patterns of user's behavior quantitatively. Some results are shown in Figure 8 (see also Ref. [19]). The temporal pattern of user activity related to the set of popular posts considered in this work is shown in Figure 8a. The index indicating a user along the vertical axis is given according to the time of user's first appearance in the considered dataset. The times of user's actions are marked by points in the direction of time axis. Notably, some users are much more active than the others, which leads to the fractal pattern. A part of the pattern, containing 100 very active users, is analyzed in Figure 8b, where each point represents user activity within a time bin of one day. The color

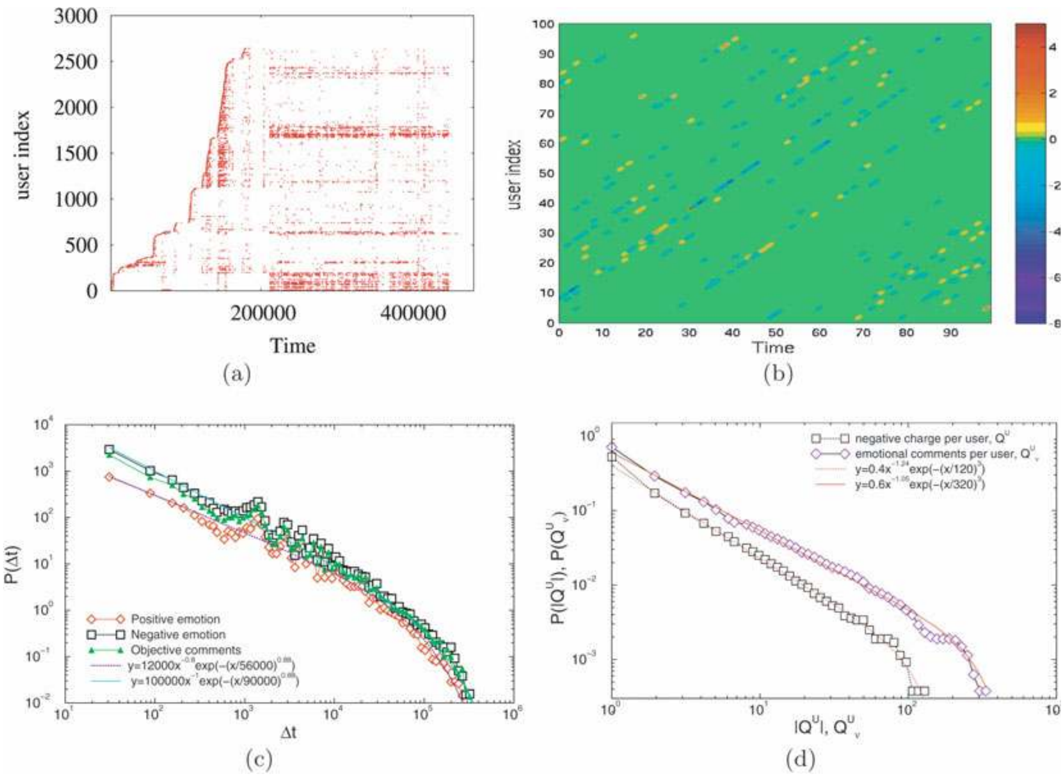

Fig. 8. (Color online) (a) Temporal pattern of user activity for the users at popular posts. (b) Part of the pattern containing 100 users is shown with color-code representing the emotional charge Q^U of the comments made by the same user within several consecutive time windows. Each time window corresponds to one day. (c) Distribution of time intervals between two consecutive positive (negative, objective) comments of a user, averaged over $N_U = 3000$ most active users. (d) Cumulative distributions of the number of emotional comments by a user, Q^U_ν, and of the negative charge of a user, Q^U, averaged over $N_U = 3000$ most active users in the dataset. The fit lines according to equation (11) are explained in the text.

code represents the emotional charge Q^U of the comments made by a particular user within a given day.

From the dataset we selected a group of $N_U = 3000$ most active users and analyzed their activity over time. For every user we calculated the time intervals Δt between two successive positive (negative, objective) comments made by a particular user at anyone of the posts. The distributions of these time intervals $P(\Delta t)$ averaged over all N_U most active users are given Figure 8c. The distributions exhibit power-law decay with stretched-exponential cut-offs, according to the general expression

$$P(X) = Ax^{-\tau_X} \exp\left(-(X/\Delta_X)^\sigma\right), \qquad (11)$$

where the stretching exponent $\sigma = 0.88$ for both positive and negative emotion comments, whereas the slopes $\tau_X \equiv \tau_+ = 0.8$, for positive, and $\tau_X \equiv \tau_- = 1.0$, for negative comments, respectively. These distributions suggest that larger delay time is more probable for positive than for negative comments. In this dataset the distribution of objective (i.e., neutral) comment closely follows the distribution of negative comments. The occurrence of the power-law dependences in the distributions like $P(\Delta t)$ was observed at different Web portals (see also [49]), whereas the slope of the distributions is characteristic for user communities at a given Blog site. In view of bipartitivity of our networks, it is interesting to mention that another robust distribution of time-delayed actions was observed at posts. Namely, the distribution of action times relative to the time of posting, $P(t - t_P)$ averaged over all posts was found to have a power-law decay for Blogs [19]. Similar distributions are observed in relation to human actions to various types of alerts and the origin of power-law behaviors discussed in reference [26,50]. Here we further focus on user's behavior only. Analysis of comments by the subset of very active users on popular posts indicates power-law decay of the distribution of the number of emotional comments made by a user within a given time period, Q^U_ν, and the absolute values of negative charge of these comments, Q^U, as shown in Figure 8d. The occurrence of the power-law decay in these two distributions suggests that a small number of users write many emotional comments (within a fixed time window). Moreover, a small number among them writes comments with large negative charge. Both distributions obey the functional form in equation (11) with $\sigma = 3$ and different slopes: $\tau_Q = 2.24$, for the charge, and $\tau_\nu = 2.05$, for the number of emotional comments, respectively. The cut-off lengths are $\Delta_Q \sim 120$, while $\Delta_\nu \sim 320$, supporting the above conclusions.

An example of the time series of user activity with marked polarity of the comments is shown in Figure 9a, indicating that although the polarity of the comments by the same user often flips from positive to negative and

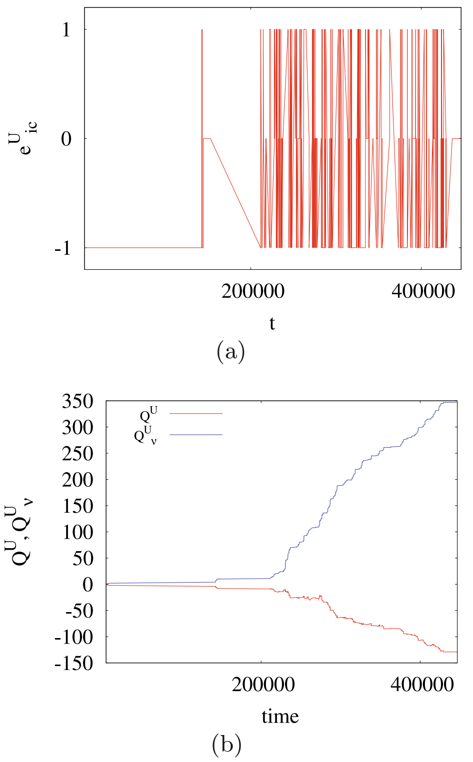

Fig. 9. (Color online) An example of a very active user time series of comments with their emotional polarity (a), and the evolution of the number of emotional comments of any polarity, Q^U_ν, and their charge, Q^U, (b).

back, the excess of the negative comments is found over the periods of user's intensive activity. This explains the appearance of the negative charge over larger time period, as shown in Figure 9b for the same user.

The actions of many users at the group of popular posts in our dataset appear to have long-range temporal correlations, which is manifested by the fractal nature of the time-series of the number of comments $N_c(t)$. In Figure 10 we show the time series of those comments which are classified as subjective by our emotion classifier. Specifically, the comments that are classified as negative are shown separately from positive comments. The power spectra of these time series appear to have $1/f$-type correlations beyond certain frequency, while the correlations vanish in the high-frequency range. Apart from the abundance of the negative comments, practically no differences can be detected in the spectrum of positive and negative comments. The periodicities superimposed to the spectrum occur, corresponding to the natural cycles (days, weeks, etc.). The occurrence of the $1/f$-noise indicates the bursting events in user behavior at Blogs. Detailed analysis of the avalanches of the emotional comments will be reported elsewhere [51].

5 Conclusions and discussion

We have demonstrated that user communities occurring in relation to popular Blogs can be readily identified and their structure and evolution analysed from the high-resolution Blog data. Due to the specific structure of the *posts-mediated user interactions*, the raise and fall of these communities is related to the contents of the posts. Specifically, we have shown that the *emotional contents* of the texts of popular posts and comments is tightly correlated with the number of users and their action over time. This is in contrast to the subject preference, which is observed at normally popular posts and some other cybercommunities [19,24,28]. We have presented suitable methodology for a systematic study of the user collective behavior at popular Blogs, which comprises of two integrated parts with:

- *bipartite network mapping*, which enables use of the theory of complex networks, in particular their topology analysis and detecting the mesoscopic community structure;
- *machine-learning methods of text analysis* for the emotion classification of the texts of comments related with the identified communities.

Within this methodology the user grouping around certain popular posts (and with them related comments) can be identified as dynamical communities in the cyberspace and their evolution followed over time. Furthermore, we find that the fluctuations in size of the community and their activity over time are tightly linked with the amount of emotions expressed in the related comments.

The occurrence of cybercommunities has a signature of the collective dynamical phenomena, which can be characterized by several quantities exhibiting power-law distributions and robust patterns of behavior, both at user and posts partitions (cf. Fig. 8 and Refs. [19,26]). These collective dynamical effects can be further studied in terms of the long-range correlations in the emotional time series, as an example shown in Figure 10, and identifying the bursting events and avalanches of the emotional comments over the networks [49].

Understanding of the mechanisms behind the observed complex dynamical behavior in cybercommunities requires theoretical modeling, e.g., by the agent-based models and the evolution dynamics. Currently the models of user emotional behavior in cyberspace are being developed both at the level of an individual agent (user) in a cyberenvironment [52,53], as well as the network-automata models [51], which take into account the basic evolution rules that are leading to the observed emergent dynamical states.

Other aspects of the emotion analysis, including multidimensional character of the emotion and their respective contributions of the fine structure of the emotional contents to the bursting events in the cyberspace, remains an open question to be studied in future.

The methodology presented here can be used for high-resolution data from other Web portals with text-based

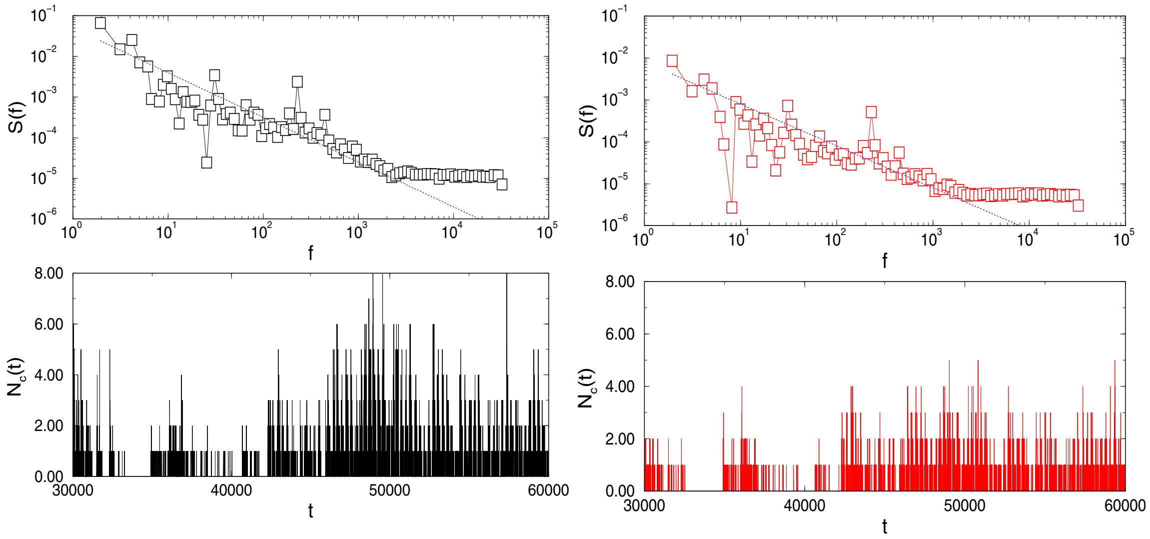

Fig. 10. (Color online) Example of the time series of the number of emotional comments $N_c(t)$ on the popular Blogs and their power spectrum $S(f)$: separated are comments which are classified as negative (left panels) and positive comments (right panels). Dotted lines indicate slopes −1.1 and −1.0, respectively.

communications and the emotion classifier appropriately trained for that data type.

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7-ICT-2008-3 under grant agreement n^o231323 (CYBER-EMOTIONS). M.M. also thanks for the partial support by COST-MP0801 action. B.T. thanks for the support from the national research program P1-0044 (Slovenia).

References

1. T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, J. Weitzner, Science **313**, 769 (2006)
2. D. Donato, S. Leonardi, S. Millozzi, P. Tsaparas, J. Phys. A **41**, 224017 (2008)
3. B. Tadić, Physica A **293**, 273 (2001)
4. J. Kleinberg, Communications of the ACM **51**, 66 (2008)
5. A. Cho, Science **325**, 406 (2009)
6. L. Adrianson, Computers in Human Behavior **17**, 71 (2001)
7. C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, V. Loreto, PNAS **106**, 10511 (2009)
8. M. Thelwall, A. Byrne, M. Goody, Information Research **12**, 327 (2007)
9. G. Brumfiel, Nature **459**, 1050 (2009)
10. T. Zhou, H.A.T. Kiet, B.J. Kim, B.H. Wang, P. Holme, Europhys. Lett. **82**, 28002 (2008)
11. D. Derks, A.H. Fischer, A.E.R. Bos, Comput. Hum. Behav. **24**, 766 (2008)
12. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, J. Am. Soc. Inf. Sci. Technol. (in press), 2010
13. J.H. Fowler, N.A. Christakis, British Medicine Journal **337**, a2338 (2008)
14. J. Bollen, A. Pepe, H. Mao, `arXiv:0911.1583`, 2009
15. F. Fu, L. Liu, K. Yang, L. Wang, The structure of self-organized blogosphere, `arXiv:0607361`, 2006
16. L. Liu, F. Fu, L. Wang, Information propagation and collective consensus in blogosphere: a game theoretical approach, in *Web 2.0 - Eine empirische Bestandsaufnahme* (Vieweg+Teubner Verlag, 2008), pp. 87–104
17. J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst, Cascading Behavior in Large Blog Graphs, in *Proceedings of 7th SIAM International Conference on Data Mining (SDM)* (2007), p. 2940613
18. Y. Sano, M. Takayasu, *Macroscopic and microscopic statistical properties observed in blog entries*, J. Economic Interaction and Coordination (2010)
19. M. Mitrović, B. Tadić, Eur. Phys. J. B **73**, 293 (2009)
20. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, Phys. Rep. **424**, 175 (2006)
21. B. Tadić, G.J. Rodgers, S. Thurner, Int. J. Bifurc. Chaos **17**, 2363 (2007)
22. J. Grujić, Lect. Notes Comput. Sci. **5102**, 576 (2008)
23. J. Lorenz, Eur. Phys. J. B **71**, 251 (2008)
24. R. Lambiotte, M. Ausloos, Phys. Rev. E **72**, 066107 (2005)
25. R. Lambiotte, M. Ausloos, Eur. Phys. J. B **50**, 183 (2006)
26. R. Crane, F. Schweitzer, D. Sornette, Phys. Rev. E **81**, 056101 (2010)
27. P. Panzarasa, T. Opsahl, K. Carley, J. Am. Soc. Inf. Sci. Technol. **60**, 911 (2009)
28. J. Grujić, M. Mitrović, B. Tadić, IEEE Xplore, 259 (2009)
29. T. Zhou, L. Jiang, R. Su, Y. Zhang, Europhys. Lett. **81**, 58004 (2008)
30. B. Kujawski, J. Holyst, G. Rodgers, Phys. Rev. E **76**, 036103 (2007)
31. L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group formation in large social networks: membership, growth, and evolution, in *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2006), pp. 44–54
32. B. Fortuna, M. Grobelnik, D. Mladenic, Ontogen, `http://ontogen.ijs.si/`, 2008
33. P. Dodds, C. Danforth, Journal of Happiness Studies, 1389 (2009), ISSN 1389-4978 (Print) 1573-7780 (Online)
34. B. Pang, L. Lee, *Opinion Mining and Sentiment Analysis* (Now Publishers Inc, 2008)

35. M. Mitrović, B. Tadić, Phys. Rev. E **80**, 026123 (2009)
36. V. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, J. Stat. Mech. **2008**, P10008 (2008)
37. S. Fortunato, Phys. Rep. **486**, 75 (2010)
38. B. Bollobas, *Modern Graph Theory* (Springer, 1998)
39. C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, 1st edn. (Cambridge University Press, 2008)
40. C.D. Manning, H. Schüetze, *Foundations of Statistical Natural Language Processing*, 1st edn. (The MIT Press, 1999)
41. I.H. Witten, T.C. Bell, IEEE Transactions on Information Theory **37**, 1085 (1991)
42. C. Macdonald, I. Ounis, *The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection*, Technical report, Department of Computing Science, University of Glasgow, 2006
43. C. Macdonald, I. Ounis, I. Soboroff, Overview of the TREC-2008 blog track, in *The Sixteenth Text REtrieval Conference (TREC 2008) Proceedings* (2008)
44. A. Lancichinetti, S. Fortunato, Phys. Rev. E **80**, 056117 (2009)
45. M. Mitrović, B. Tadić, Lect. Notes Comput. Sci. **5102**, 551 (2008)
46. M. Rosvall, C. Bergstrom, PNAS **105**, 1118 (2008)
47. A.N. Samukhin, S.N. Dorogovtsev, J.F.F. Mendes, Phys. Rev. E **77**, 036115 (2008)
48. L. Donetti, M.A. Muñoz, J. Stat. Mech.: Theory and Experiment **10**, P10012 (2004)
49. M. Mitrović, G. Paltoglou, B. Tadić, *Quantitative analysis of bloggers colelctive behavior powered by emotions*, in preparation, 2010
50. G. Grinstein, R. Linsker, Phys. Rev. E **77**, 012101 (2008)
51. M. Mitrović, B. Tadić, Network automaton model of bursting emotional behavior on Blogs, preprint (2010)
52. D. Garcia, F. Schweitzer, *Emotions in product reviews – Empirics and models*, in preparation, 2010
53. D. Garcia, F. Schweitzer, Eur. Phys. J. B **77**, 533 (2010)