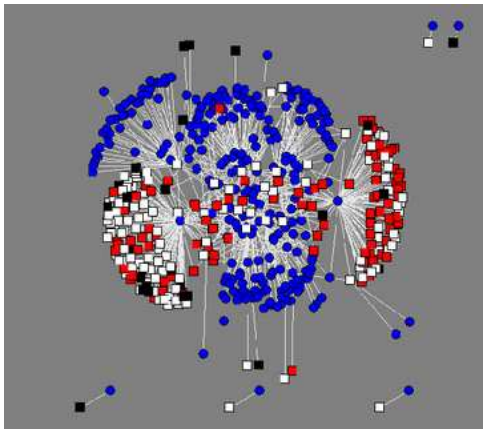
CYBEREMOTIONS - Collective Emotions in Cyberspace
EU FP7 funded IP project, Grant agreement no. 231323

CYBER**EMOTIONS** project
Workpackage 6– Jožef Stefan Institute - Department of theoretical physics

Deliverable "D6.1: *Estimate of the conditions for the self-organized critical emotional states in e-communities based on realistic parameters and their robustness with respect to parameter variations*"

From February 1*st*, 2009 **to January** 31*st*, 2010



February 15, 2010

CYBER**EMOTIONS** coordinator: Prof. Dr. Janusz A. **Holyst**

WP6 leader: Prof. Dr. BOSILJKA **TADIĆ**

WP6 other team members:
Ms. MARIJA **MITROVIĆ**
Dr. MILOVAN **ŠUVAKOV**

Authors: TADIĆ B. and MITROVIĆ M.

Contents

1

Abstract

We analyze the high-resolution data from Blogs and Diggs and dialogs between friends along the MySpace network, which contain information about user actions and full text of their comments posted over time. The data are mapped onto bipartite networks with users, as one partition, and posts and comments, as another partition, on which users interact via posted material. With the machine-learning methods the text of the posts and comments is analyzed against emotional contents, classified as positive, negative, or objective. Within the framework of physics of complex dynamical systems and network theory we perform quantitative analysis of user behavior and unravel the role of emotions in the collective phenomena emerging within the user communities, which are grouped around popular posts. We give the conclusive evidence that the communities self-organize into dynamical critical states with avalanches of the emotional comments lasting over large periods of time, and determine the control parameters which have the potential to tune the emergent states.

2

Deliverable D6.1: Objectives & Research Methodology

Deliverable D6.1: Estimate of the conditions for the self-organized critical emotional states in e-communities based on realistic parameters, and their robustness with respect to parameter variations.

2.1 Objectives of D6.1

Mapping the high-temporal resolution data which include information about Users and their Posts and Comments onto bipartite graphs. Then using the graph theory methods and advanced statistical analysis suitable for complex dynamical systems and classified emotional contents of the related texts, give quantitative evidences of the self-organized critical states with robust patterns of user behaviors and extract related control parameters which can tune the emergent states.

2.2 Research approach and methods used

The complex task of the data-driven modeling and extracting the relevant parameters required suitable approaches and number of numerical methods, which we have developed, summarized as the following sequence of steps:

- Data collection by developing Python scripts (Blogs, MySpace networking);
- Mapping the large data streams onto bipartite networks with [Users and Post+Comments] partitions;
- Topological analysis of these bipartite networks and their projections; setting additional filtering criteria for the data according to the network topology measures;
- Spectral analysis of the emergent networks and Detecting the User-Communities and the related Posts;
- Running the Emotional Classifier (developed by WP3) for the emotional contents in all Post+Comments of the selected user communities; analysis on the subgraphs containing the emotional comments;
- Analysis of temporal patterns of user behaviors and tracing the emotional avalanches over the networks;
- Defining and extracting the realistic parameters behind the observed behaviors;
- Theoretical modeling (Cellular Automata Models on bipartite networks) of the emotional avalanche dynamics within the real networks structure and parameters to test robustness of the observed emotional states.
- Graphical visualization of the bipartite and projected networks, and the emotional avalanches on them.

2.3 Data sources and structure

For our *network approach to techno-social interactions* we need the high-resolution data of the temporal occurrence of the events (with **time resolution of** 1 **minute**) and full information about both **Users** and **Posts-and-Comments**, as well as the full **Text** of the Post and Comments. Using the bipartite network approach and the temporal behavior of users over these networks, we have done full range of analysis, listed above, on several sets of data on Blogs and Diggs, and of MySpace-networks data. These data have been collected by our group (Blogs B92 and BBC Blogs, and MySpace networking data), and by Wolverhampton group (Diggs data). Depending on the source (policy of the Website), several variations may appear regarding the information available in the datasets, which may limit the output networks. Specifically, the data that we considered are described here:

- **Blogs data** are collected by us starting from March 2009. We have used two Blogsites, B92 Blogs and BBC Blogs, with entirely different structure and managing policy. We were able to download and analyse the data from B92 Blogs (Belgrade radio B92) from its opening in 2007 until March 2009, where the policy with completely free access, authorship and subject structure of Posts was practiced. This enabled us to study self-organizing dynamics and the emergence of communities and groupings of Posts according to how the Users considered them, as we reported in Ref. [15]. We also downloaded the data from BBC Blogs from the same period (size of all data are given in Table 2.1). In the case of BBC Blogs the Post are pre-classified by Categories, authorship of Posts is not free, and information about Comment-on-Comment is not available.

- **Diggs data** have been collected by WP3, and made available to us in September 2009. The data contain full information as we need it, including User IDs, Post IDs and Comment-on-Comment information and emotion classification, done by WP3, with binary values ± 1 for the emotional, and else objective text;

- **MySpace-networked data** are small portions of the huge network of "friends" from MySpace, collected by our robot around different locations and time windows (see details in Chapter devoted to networks). The structure of MySpace is highly inhomogeneous and accessibility of the networked data together with dialogs between connected nodes appeared as a major problem. Recently we managed to make a robot which is able to adapt to many of potential pitfalls along the crawling for the network of connected users and have done some preliminary analysis of the collected networks.

Table 2.1: Summary of the data analysed by IJS group (WP6). [Abbreviations: U=user, P=Post, C=Comment;]

Name	Source	Collector	Script/Author	Emotional content	Data size	Analysis
BBC Blog	www.bbc.co.uk /blogs/	JSI group	python script M. Mitrović	IJS group done emotion analysis with Classifier by UW group binary	$N_P = 3972$ $N_C = 80873$ $N_U = 21462$	time patterns; U+P+C networks; statistics of emotions within user communities; emot.avalanches;
B92 Blog	blog. b92.net	JSI group	python script M. Mitrović	in Serbian no classifier!	$N_P = 4784$ $N_C = 406527$ $N_U = 4598$	statistics; U+P+C networks; time.patterns; user communities;
My. Space	www. myspace. com	JSI group	python script M. Šuvakov	text cleaning! -	$N_U = 64738$ $N_c = 172128$	network structure; time-series; time.windows;
Digg	www. digg.com	UW group	?	UW group binary	$N_P = 1195808$ $N_C = 1646153$ $N_U = 484986$	time.series; statistics; networks U+P+C; user communities; emot.avalanches;

3

Data to Networks & Finding User Communities

3.1 MySpace networks

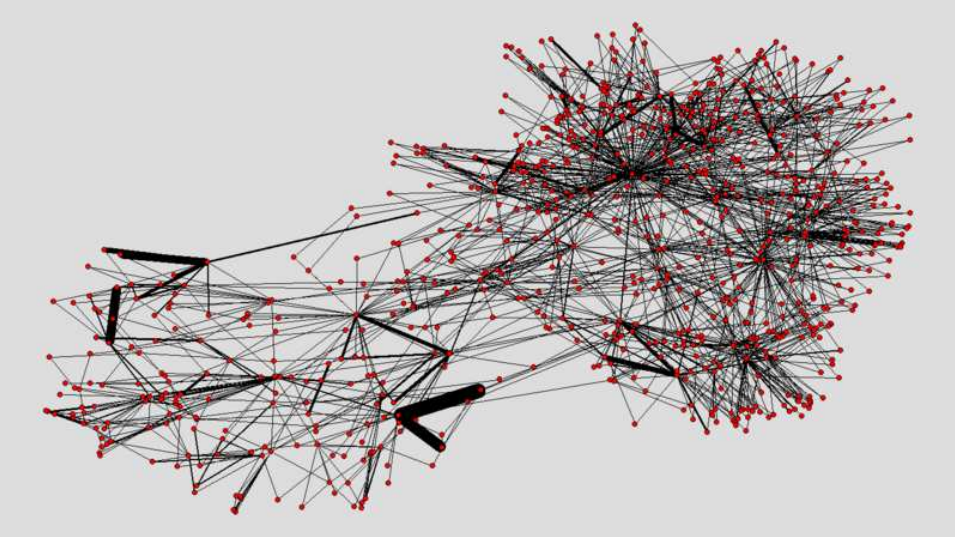


Figure 3.1: An example of MySpace network collected by our robot. Nodes represent Users, who are connected by the dialogs (indicated by the widths of Links) occurring within a specified time window.

We have developed a robot to collect the data from MySpace in *terms of a connected network and related dialogs*. The script works in a kind of *first breath* search starting from a specified position (user) and obeying given conditions. The search is made versatile to avoid a large number of pitfalls (e.g., due to diversity of user profiles and missing or non-public data stored). More importantly, for the crawling we can change several parameters, depending on the situation found at a site or according to specified requirements:

- initial position, depending on the user properties, number of connections and recent history of the activity;
- time depth, or time-window T_{WIN} of the dialogs (typically 2 months, 3 months, or more);
- topological depth d_i, i.e., number of "breaths" away from the starting position at the node i;

All these parameters affect the outcome network of users (an example is shown in Fig. 3.1). During the crawling we collect all data about the users with public profile, and all texts of the dialogs occurring within a given time window that we are interested in. The data are automatically stored in an SQL database. Up to now we have collected several datasets which contain networks with up to $N_U \approx 6.4 \times 10^4$ connected users and $N_c \approx 1.7 \times 10^5$ their dialogs in two different time windows. Some topological features of such networks are shown in Fig. 3.2, indicating also the time-window effects on the emergent network topology. These are recent developments in our research of MySpace networks and full analysis of this type of data is still to be conducted.

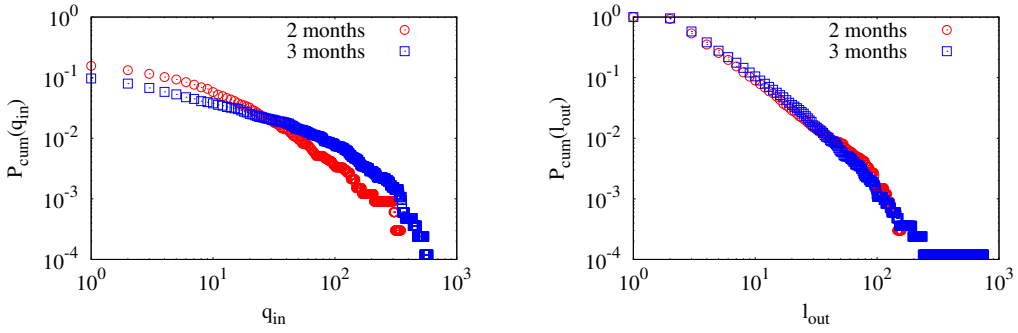


Figure 3.2: For MySpace networks collected starting from the same position and including links which are effective within two different time windows: Cumulative distributions of (a) in-degree, and (b) out-strengths of nodes (users).

3.2 Bipartite networks of Blogs and Diggs

The data from Blogs and Diggs, which we analyse in all details here, are of different structure: we have high resolution in time of the data on both Users and Posts and Comments, which we map onto a *bipartite network*. On these networks Users are recognized as one natural partition, and Posts and Comments together, as another natural partition. This representation is particularly suitable for the users in the Cyberspace of Blogs (and similar Web portals, see also [8, 12]), since in such networks a link is allowed only between different partitions! Depending on the type of user activity, i.e., reading the existing text or posting a new text, we distinguish the directions of the links as follows: Post (Comment)→User represents *the User reading the Post (Comment)*, while User→Post (Comemnt) indicates that *the User is writing a new Post (Comment)*.

In our research within the Project we use several types of networks derived from the data:
(a) *Full bipartite network*, as described above, with Users (U) and Posts & Comments (P&C), which is most suitable for visually presenting a *network of an individual Post with all its Users and their Comments*. Notice that these networks are *directed* graphs, which provides additional features of the dynamical system. In the considered datasets such network can be very large (see Table 2.1 for the size of data);
(b) *Weighted bipartite network* is a compressed bipartite network with Users and Posts as nodes (U+P), while the widths of links W_{ij} represent the number of comments of the user i on the Post j. These networks are undirected and more suitable for the spectral analysis, compared to full bipartite networks. An example of such network representing the evolving Community on Blogs is given in Fig. 3.3 (detailed analysis i given in [13]);
(c) *Weighted networks in User projection or Post projection*, are undirected monopartite versions obtained by suitable projections from the full bipartite network of the data. In the projection two Users are connected directly with weighted links, representing number of common Posts per pair of Users, C^B_{ij}, or the number of common Users per pair of Posts, C^U_{ij}, in the Post projection (see Refs. [15, 8]).
We have explored in all details the topological properties of the networks derived from the data of Blogs and Diggs [15, 13]. For the purpose of this report, it is useful to mention that the topology analysis alone already suggest a clear Post popularity measure (number of comments exceeding certain value, here ~ 100, which is visible as a break point in the degree distributions and the distributions of commons in each considered dataset). The popular

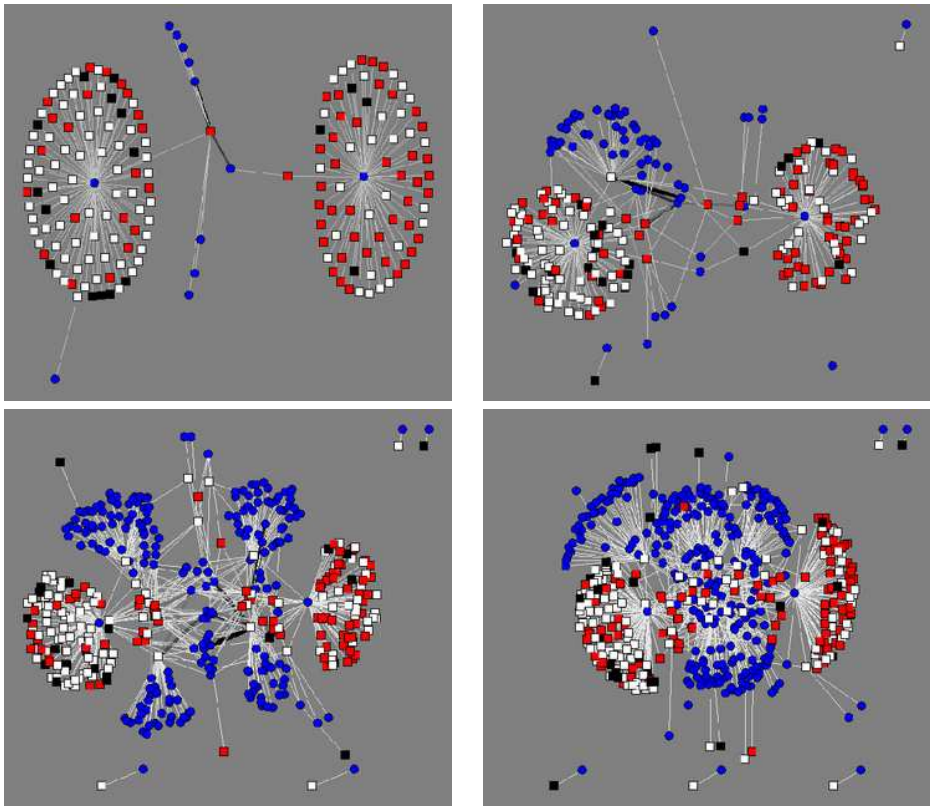


Figure 3.3: Evolution of a weighted network on Blogs of normal popularity: Users (blue circles) linked to Posts (squares), which are colored by their net emotional content: positive (red), negative (black), or objective (white).

Posts appear to have different power-laws. We concentrate mainly on popular Posts as most suitable for detecting the collective emotional behavior. Moreover, a striking topological property of these networks is their *mesoscopic* inhomogeneity, or the occurrence of *communities* of nodes with stronger connections among each other [7, 14]. These communities play an important role in the collective dynamics that we are interested in and thus we have developed a systematic way to detect them. The methodology based on the *eigenvalue spectral analysis of the weighted bipartite graphs* is outlined below (detailed study of modular monopartite networks is given in Ref. [14, 15]).

3.3 Identification of user communities on Blogs and Diggs

Having the network constructed from the data and put into a suitable form (as discussed above), we look for the community structure on that network by the eigenvalue spectral analysis of the Laplacian operator, which is related to the network adjacency matrix. In particular, for the networks of popular Posts from Blogs and Diggs, we find that the appropriate forms of the networks [15] are the weighted bipartite networks of Users and Posts, for which the normalized Laplacian is given by [5]

$$L^U_{ij} = \delta_{ij} - \frac{W_{ij}}{\sqrt{\ell_i \ell_j}}\,; \qquad \ell_i \equiv \sum_j W_{ij}\,. \tag{3.1}$$

Here ℓ_i is so called *strength* of a node (that can be defined both for User or Post nodes), and W_{ij} are the elements of a real symmetrical matrix defined by the widths of the links, which are uniquely determined as the number of comments of the User i to the Post j. In the case of User-projected networks, which is more suitable for the community analysis in normally popular Blogs, discussed in Ref. [15], the matrix W_{ij} is replaced by the

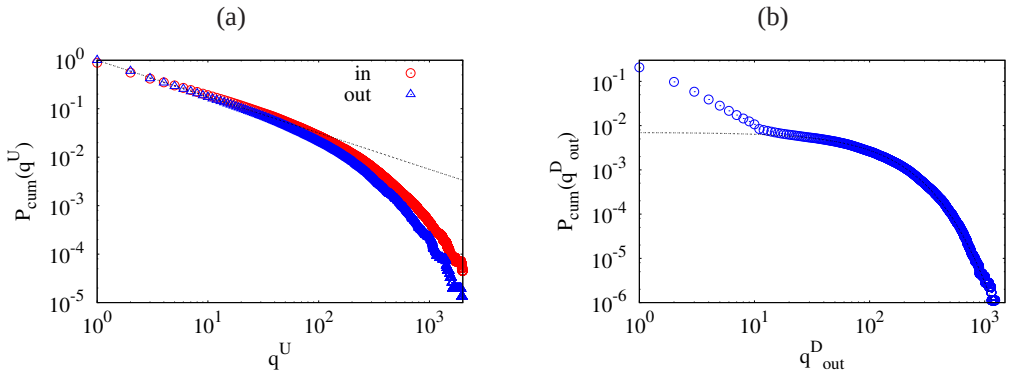


Figure 3.4: For Diggs network: in- and out-degree for user partition (a) and out-degree of post partition (b).

commons C^B_{ij}. The standard algorithms for solving the eigenvalue problem are then used to determine the complete spectrum of the eigenvalues $\{\lambda_0, \lambda_1, \cdots \lambda_N\}$ and the corresponding eigenvectors. The size of the network N can be sometimes a limiting factor for the numerical algorithm. In such cases we reduce the network by using the topological properties, e.g., node's connectivity, centrality or strengths.

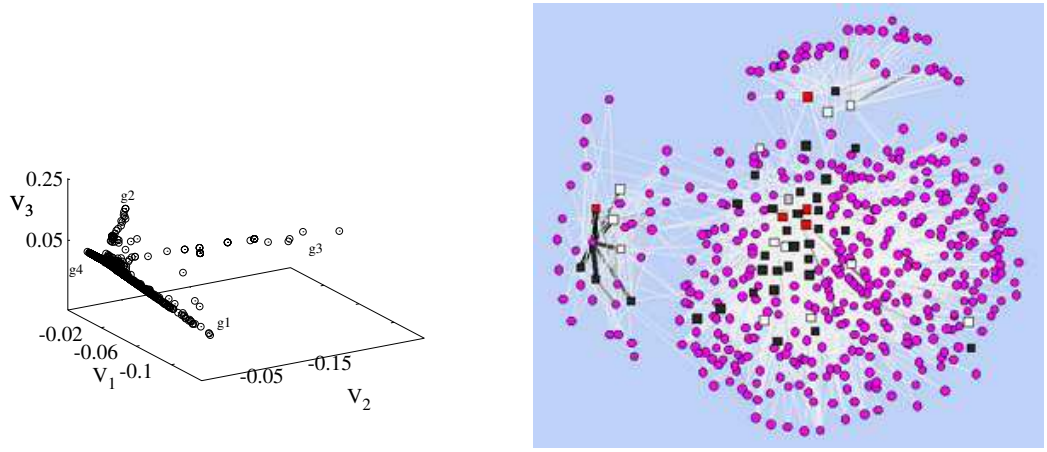


Figure 3.5: Left: Scatter plot in the space of of three eigenvectors indicating the occurrence of user communities, and Right: The community structure in the weighted network of BBC popular Blogs with Users (circles) and Posts (squares). Weights on links correspond to the number of comments posted by the User on the Post, while the color on the Posts indicates their overall emotional content, averaged over all comments on it.

Detecting the communities on such networks is based on well known properties of the eigenvalues and corresponding eigenvectors of the adjacency matrix and similarly of other matrices related to it, such as the Laplacian (3.1). In particular, in the presence of subgraphs, apart from the zero eigenvalue corresponding to the entire graph, the eigenvalue spectrum shows *few small non-zero eigenvalues*, corresponding to each of the subgraphs. These eigenvalues appear to be separated from the main part of the spectrum for a particular network realization, or a separate peak appears in the spectral-density, when an ensemble of the networks can be considered [14]. The eigenvectors belonging to these eigenvalues, are orthogonal to the eigenvector of the zero eigenvalue (which has all positive components), and hence appear to *localize* on the subgraphs. This means that non-zero components of such an eigenvector carry indexes of the nodes in a subgraph. Consequently, one can identify the nodes of each subgraph by looking at different branches in the scatterplot of the eigenvectors against each other. An example relevant for our discussion of popular Blogs is shown in Fig. 3.5 indicating four communities. Three such communities of Users with related Posts are shown on the network. Each Post is colored to indicate its emotional content: positive/negative (red/black) if the average emotion of all comments exceeds ± 0.25, else neutral (white).

4

User Behavior & Emotional Avalanches on Blogs and Diggs

4.1 Robust patterns of user collective activities

From the high-resolution data that we have one can easily extract the events at a particular post and, similarly, the activities of a particular user over time. The statistical analysis of such data [15], together with the emotional contents of the related comments [13], gives insight into the underlying mechanisms driving the activity on Blogs and Diggs. Here we mention (see also discussion below) two robust features related to the lifetime of Posts: (i) *Delay $t - t_0$ of the activity to posted material*, the distribution $P(t - t_0)$ is shown in Fig. 4.1a, exhibits robust power-law tail (data are from two Blogsites, see also different systems studied in [2]); (ii) *Number of comments $N_{com}(i)$ on a Post* over time has a particular pattern of fluctuations (inset to Fig. 4.1b). Considering all posts in the dataset we plot the dispersion of such time series against its average, which is shown in Fig. 4.1b, where each point represents one Post. The occurrence of the Taylor scaling [6] with two distinct sets of Posts is visible (corresponding to two slopes on the plot), one with predominantly *internally driven* dynamics, slope ~ 0.5, and the other set where external driving is more prominent, slope increasing towards 1. In order to further examine

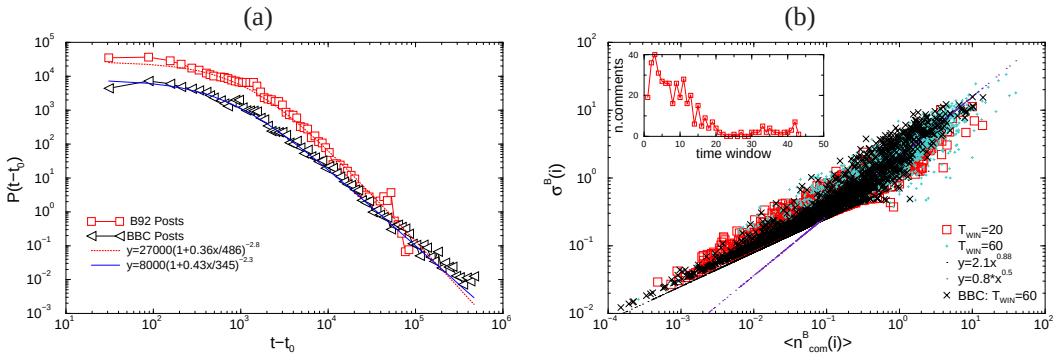


Figure 4.1: (a) Deleyed action of events at posted material on Blogs and (b) Dispersion of time-series of all Posts (different time windows are considered).

the nature of the activity on these Posts, we need to consider the network connecting them *via active users*. In particular, we find the *emergent community of users* related to specific sets of Posts and consider detailed activity of these users and their emotional comments. To study the *collective user behavior over networks* and *role of emotions* in it, we select from the Diggs dataset those popular Posts which exhibit over 50% of Comment-on-comment actions (discussion-driven or ddDiggs).

4.2 Evidence for self-organized criticality in the emotional comments

Represented as a network, popular discussion-driven Diggs make a large community consisting of $N_U = 81925$ **Users, who made** $N_C = 918019$ **Comments on** $N_P = 3984$ **Posts**. Time sequence of the cumulative activity, i.e., number of comments within a small time bin $t_{bin} = 5$ minutes on these Posts is followed within entire time period available in the dataset. Similarly, the time-series of the number of *emotional comments* $N_{ecomm}(t)$ and number of *negative-emotion comments* within the time bins is shown in Fig. 4.2. (Only a part of the time-series is shown. The community exists for 3229 hours, of which the intensive activity with avalanches occurs over 2153 hours.) Two characteristic features are immediately recognizable: (i) The time-series are fractal (antipersistent), with the long-range correlations of the $1/\nu$-type over a large frequency range (except for very large frequency); (ii) Each peak in the time-series, indicating an *avalanche* of comments within certain time period, is superimposed onto such a peak of *negative* comments, suggesting a particular role of negative emotions in the dynamics on Diggs.

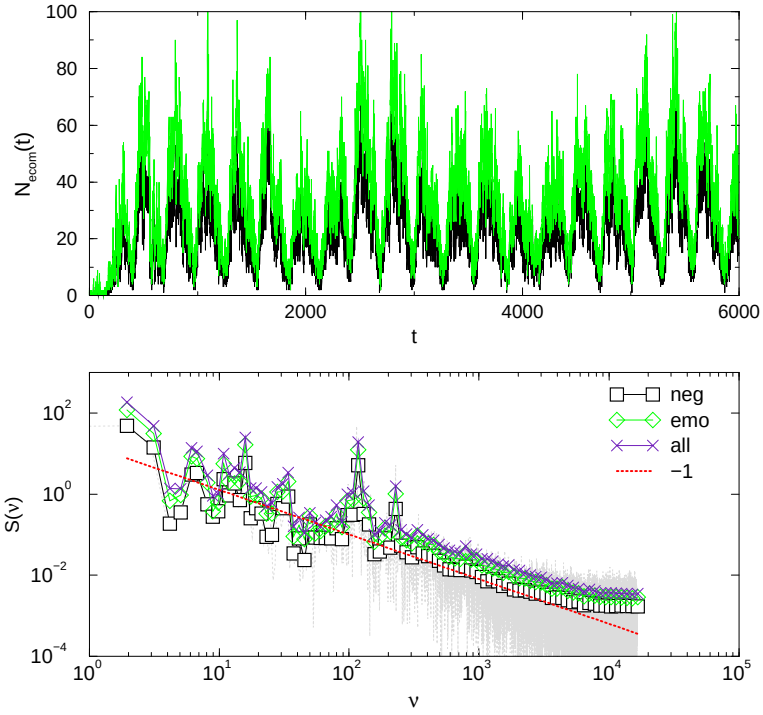


Figure 4.2: Timeseries showing the cumulative emotional avalanches in large communities at popular ddDiggs (top); Power-spectrum (bottom) shows long-range correlations with $1/\nu$ type of noise occurring in the timeseries of negative comments, all subjective (emotional) comments, and all comments including objective ones.

Occurrence of the temporal correlations, manifested in the *fractality* of the time-series and the power-spectrum of $1/\nu$-type, is an indication of the *self-organizing criticality* (SOC) in this dynamical system. Further evidences of the SOC [3, 9, 19, 4, 16] can be found, typically in power-law distribution of the avalanche sizes according to the expression

$$P(s) \sim s^{-\tau_s} exp(-s/s_0)\ ; \tag{4.1}$$

where s_0 is a finite-size cut-off, and similarly for the avalanche durations $P(T)$, and other related quantities [4, 18, 17]. Furthermore, the distribution of temporal distance between consecutive avalanches, $P(\delta T)$, also exhibits a broad power-law behavior, as seen for instance in the case of earthquakes [1] and other nonlinear systems with self-organizing dynamics [10].

Here we give such evidences of the SOC by focusing first on the events occurring at each of 3984 *individual posts* in our dataset. The results for several representative quantities are given in Fig. 4.3. The time-series of the emotional comments at a particular post, Fig. 4.3a, shows two large peaks and several smaller peaks, indicating

avalanches with different number of comments on the Post. Note that an avalanche can be determined from the time series as an area under the signal between two consecutive points where the signal drops to the base line, as it is usually done in the experiments [17]. The dispersion of the signal plotted against its average value represents a single point on the plot in Fig. 4.3b. Shown are the results for all Posts in the considered dataset. Occurrence of the scaling behavior according to (see recent review in [6]) $\sigma_i \sim < N_{ecomm}(i) >_t^{\mu}$, with $\mu = 0.5$ indicated by the straight line, suggests *endogeneously* driven dynamics on these popular Posts. The size of avalanches, measures in terms of the number of comments in consecutive time bins before the activity stops, etc. along the entire time-series, is computed for all Posts in our dataset. The distribution of the avalanche sizes averaged over all Posts is given in Fig. 4.3c. Three lines are for the distributions of the size of avalanches of all comments, comments with negative, and comments with positive emotional contents. The distributions of sizes of the avalanches which contain connected negative comments (black line), and all avalanches (pink) are fitted by the expression (4.1) with the scaling exponent $\tau_s \sim 1.5$, indicating closely SOC behavior. However, the avalanches which contain the connected positive comments (red) seem to be away from the critical state, with a dominant cut-off, which suggests subcriticality. These findings are in agreement with another measure of the SOC states, the distribution of the time between avalanches, shown in Fig. 4.3d. Again, the avalanches of positive comments exhibit an exponential decay, while the avalanches of negative (and all) comments have a broad power-law distributions.

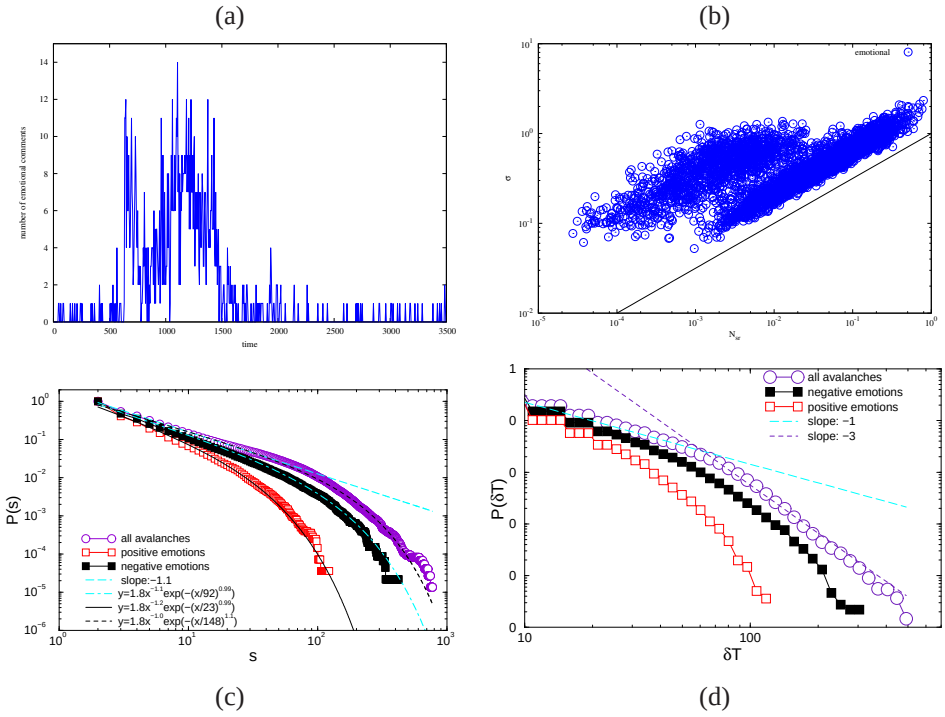


Figure 4.3: Example of the timeseries of the number of emotional comments on a popular Posts in the ddDiggs dataset (a) and the respective scatter plot of the dispersion-vs-average of the timeseries of emotional comments at all such Posts (b). Distribution of (c) the sizes and (d) the temporal distance between the emotional avalanches.

The observed differences in the distributions related to the subsets of positive- and negative-emotion avalanches of comments, as shown in Figs. 4.3c and d, suggest that there might be a control parameter, deducible from the human behavior on Posts, which can tune the emergent self-organized states in the blogging dynamics.

Our aim here is to further investigate how such SOC states occur and their robustness with respect to potential control parameters. The strategy is as follows: In the next section we extract several such parameters, which in our opinion are relevant for the microscopic dynamics. We then introduce a cellular-automaton model in which we can first reproduce the SOC dynamics of the type documented above in Figs. 4.2 and 4.3. Furthermore, within the model these control parameters are varied in a wider range around their values extracted from the real dataset, and the predictions made of their impacts onto the SOC states of the system.

5

Control Parameters & Predictions of SOC Dynamics

5.1 Extracting control parameters from real data

Based on our experience with Blogs and Diggs, the microscopic dynamics (user–posting–comment, triggering more users for their actions, etc.) depends on several facts, which can be formulated in terms of the *update rules* and *constraints* which affect the course of the dynamics, and thus the emergent states. A minimal set of such *control parameters* regulating the dynamics with the emotions is given here are related to the real data (see the related Fig. 5.1 and Table 5.1):

- *User delayed action* to posted material; data analysis gives distributions with a power-law tail $P(\Delta t)$, see Fig. 5.1a; The parameters are: the slope τ_Δ and, in some cases, the threshold (or characteristic scale) Δ_0;
- *User tendency to a negative comment*, measured by the parameter α, which can be inferred from a given set of data as a fraction of negative comments among all comments made by the same user, and averaged over all users in the considered set; The distribution $P(\alpha)$ for a set of ddDigs is given in Fig. 5.1d.
- *Post strength*, S, is a topological measure uniquely defined on our bipartite networks as a sum of all weighted links of a Post node (that is, number of users linked to it and multiplicity of their action). It takes into account the number of comments on the Post, which is a measure of attractivity (relevance) of the posted material; Post strengths from the data are given by a stretch-exponential distribution, see an example in Fig. 5.1c.
- *User dissemination activity*, parameter λ, is a measure of contingency of the bloggers activity, and can be deduced from the data as the average fraction of the actions of users who are active more than once and at different Posts within a small time bin (we set time bin $t_b = 5$ minutes); In real data this is a very small fraction of users, however, it appears that they play an important role in accelerating the activity on a Post or transmitting an activity (and possibly emotion) from one Post to another one at the same Blogsite.
- *Network structures* mapped from the real data *at various instances of time*, are important for the quantitative analysis but also play an important role in the evolution of real events; The network of users at Posts evolves over time and new users appear, however, among the already registered users those who are very active often return to the Posts that they have commented previously. The actual network structure predicts where these type of actions may occur.

In Fig. 5.1b we also show a small part of a pattern of user emotional actions over certain period of time. Colors indicate emotional charge (difference between the number of positive and the number of negative comments) by a particular user within one-day time bin. Users are ordered by time of their first appearance (within the dataset). Comments by each particular user can be traced along the time axis, whereas the diagonal stripes indicate potential chain of the emotional activity on a particular Post.

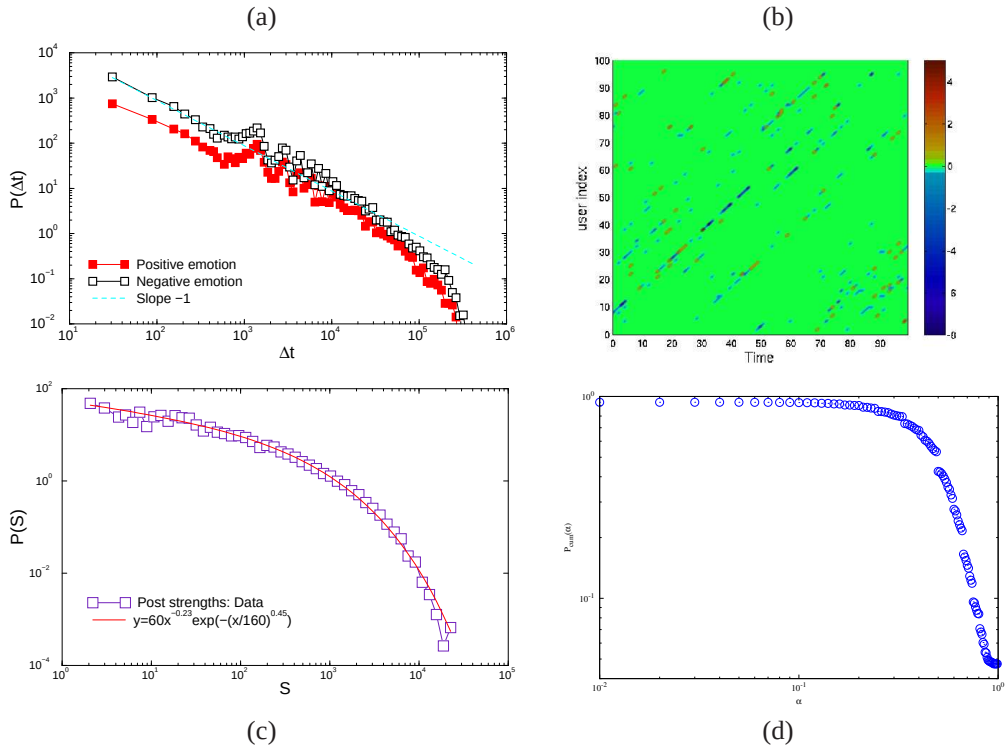


Figure 5.1: (a) Distributions $P(\Delta t)$ of the temporal delay Δt between two emotional (positive/negative) comments of the same user on BBC Blogsite, averaged over all users in the considered dataset. (b) Patterns of user emotional activity over time (for better resolution only a small part is shown) where color code indicates overall "charge" of comments made by the user within one day time bin. (c) Cumulative distribution of the strength of Posts averaged over all Posts in the set of popular Diggs. (d) Cumulative distribution of the probability α for a user to post a negative comment, averaged over all users in the dataset of popular Diggs.

5.2 Varying the parameters within CA model on realistic network

In order to grasp into the role that a particular control parameter has in the bloggers dynamics, we need a suitable theoretical concept and a numerical model, within which such parameter can be varied in a wider range compared with its value extracted from the dataset. A suitable model in the case of data and parameters studied above is a *cellular-automaton* (CA) model, defined on the bipartite network structure, which is deduced from the real data and supplemented with the following set of update rules. The procedure implemented in C++ is as follows: *(i)* Weighted bipartite network (like the one in Fig. 3.5) is constructed from the considered set of data; *(ii)* Each Post node on the network is attributed its real strength, computed from the data; Similarly, each User node is attributed its actual list of links with their weights and a probability α from the distribution $P(\alpha)$ (cf. Fig. 5.1d); *(iii)* An active User preferentially selected to start the dynamics by commenting one of the Posts from its list; With the probability α associated with the user, the comment gets negative emotions, else equal probability of being positive or neutral applies; *(iv)* Users linked to the active Post are prompted, however, each of them takes its delay time, which is taken from the distribution $P(\Delta t)$ of the actual dataset; Users whose delay-time appear to be shorter than the preset time bin t_b, are considered as active and comment the Post; Within the same time bin, with the probability λ each active user choses to comment additional Posts from its list. The cycle continues from the list of Posts being active within the preceding time bin, and so on, potentially producing an avalanche of comments.

We use time bin $t_b = 5$ minutes, same as in the analysis of the real data. Following each comment on a particular Post the available strength at that Post is reduced by one. In this way, an avalanche of connected events may stop when either none of the prompted Users chose to be active, or the Post strength is exhausted, or the network topology does not permit a link to other still active nodes. The simulations results based on the parameters and

the network structure extracted from popular Diggs data are shown in Fig. 5.2 and summarized in Table 5.1.

Fixing the set of parameters at their values extracted from the real data, and varying the parameter λ, we find the power-law distributions of the avalanche sizes for all comments and for the emotional (positive/negative) comments when $\lambda = \lambda_c \sim 2.1 \times 10^{-4}$ (see more discussion below). Varying the parameter λ in the simulations appear to have major effects on the emergent avalanche sizes. Specifically, for $\lambda < \lambda_c$ the cut-off limiting the power-law observed at λ_c dominates, indicating *subcritical* behavior. Conversely, when $\lambda > \lambda_c$, we observe excess of large avalanches, compatible with *supercritical* behavior in SOC systems. Other dynamical variables are also sampled (i.e., time-series, duration of avalanches) which are compatible with the critical state at λ_c.

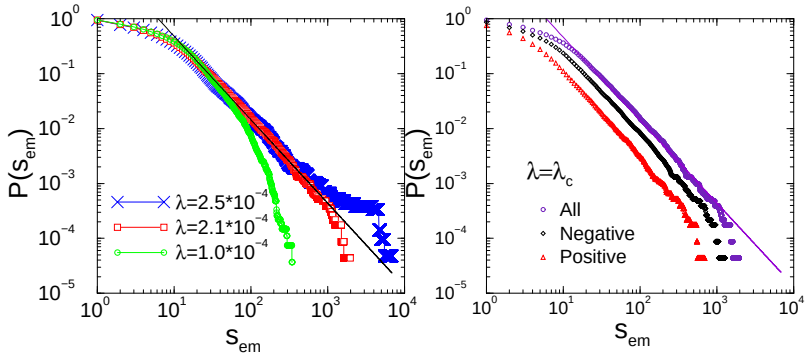


Figure 5.2: Simulated avalanches on realistic network and with control parameters extracted from the real data: Testing the role of dissemination λ (left) and the distribution of emotional avalanches in the critical state (right).

Table 5.1: Extracted from the datasets: Parameters which control the critical behavior on Blogs and Diggs.

Description/ meaning	Parameter symbol	Range in Data	Varied in CA model	Effects on the emergent states
User action delay	τ_Δ Δ_0	1.33-2.15 $< 10^3$min	real: Power-law Exponential	SOC other universality class
User emotion pref.	α A_0	$0 \leq \alpha \leq 1$ $\sim 10^{-1}$	real power-law	SOC SOC, less negative
Post relevance	S (strength)	stretch-exponential $P(S)$	real -	SOC -
User dissemination	λ (probability)	distribution! $0 \leq \lambda \leq 1$	$1.2 - 2.8$ $\times 10^{-4}$	$\lambda_c = 2.1 \times 10^{-4}$ Sub- Super-criticality
Network structure	N, CC $P(W)$	$N \gtrsim 1 \times 10^3$ power-law	real real	$\lambda_c(N)$ SOC

Varying other parameters while $\lambda = \lambda_c$ remains fixed we further test the robustness of the observed power-law avalanches. Specifically, assuming a different type of distribution for the user-delay $P(\Delta t)$, e.g., with an exponential decay (see Table 5.1), we find that the power-law avalanches still exist but with another set of the critical exponents. Similarly, assuming a power-law distribution for the user preference to the negative comments $P(\alpha)$, which is entirely different from the one derived from the real data, the SOC persists but with a reduced fraction of the negative avalanches. It is expected that the user behavior, which is modeled by the above parameters and distributions may vary, for instance by the coupling to the external environment, or by changing geographical location. Thus our simulations suggest to what direction such changes can drive the criticality at the popular Blogs and Diggs. The network structure and the strengths of Post nodes, on the other hand, are the fingerprints of the considered dataset, thus they are not varied within the simulations. Theoretically, the topological parameters of the network and of the nodes can be varied and their effects on the dynamics studied in the same manner. Specifically, we expect the value of the critical parameter λ_c to depend on the network size (number of users and comments) and node strength (posts attractivity). Hence, the network structure need to be varied according to the actual data, in order to theoretically predict the directions of the future events within the considered dataset.

In the above CA model we consider the control parameter λ as a free parameter. The actual value of this parameter can be also determined from the considered dataset. We find the *average value* of this parameter as $\lambda = 3.75 \times 10^{-4}$ from the same dataset corresponding to the time-series of the emotional comments shown in Fig. 4.2. Note that, according to our model, this value implies a *supercritical behavior* in this community, where different Posts are connected by the Users, in contrast to the case of isolated Posts, with the avalanche distributions shown in Fig. 4.3c. Computing the distribution of the avalanches in the whole community we find the excess of the large emotional and negative avalanches (see Fig. 5.3, right). Note that occurrence of such avalanches is also visible (large peaks) in the corresponding time series in Fig. 4.2. Similar analysis of the emotional avalanches was done for a large community on popular Blogs, shown in Fig. 5.3, left, where the overall behavior is compatible with the critical or sub-critical state (we found $\lambda = 1.37 \times 10^{-4}$ for this dataset).

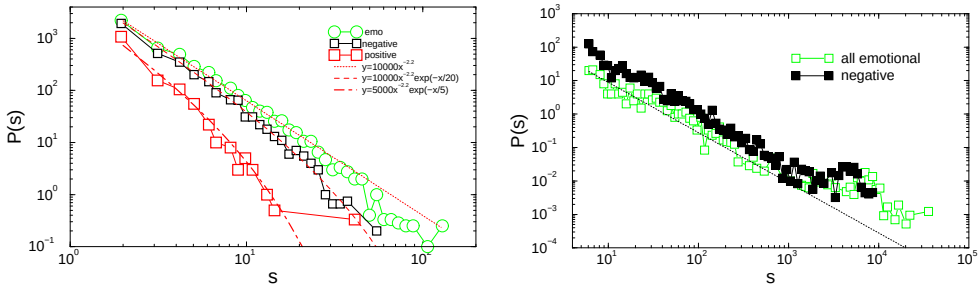


Figure 5.3: Distribution of the emotional avalanches on connected Posts in Blogs (left) and in ddDigs (right).

In the case of MySpace networks we consider the time-series of the *dialogs between users along the networks* occurring within two different time windows, shown in Fig. 5.4. We find that the avalanche-like dynamics also occurs in this case, however, with quantitatively different features. As mentioned above, the cumulative avalanches can be determined from the time-series. The distribution of the avalnche sizes is given in Fig. 5.4, indicating a

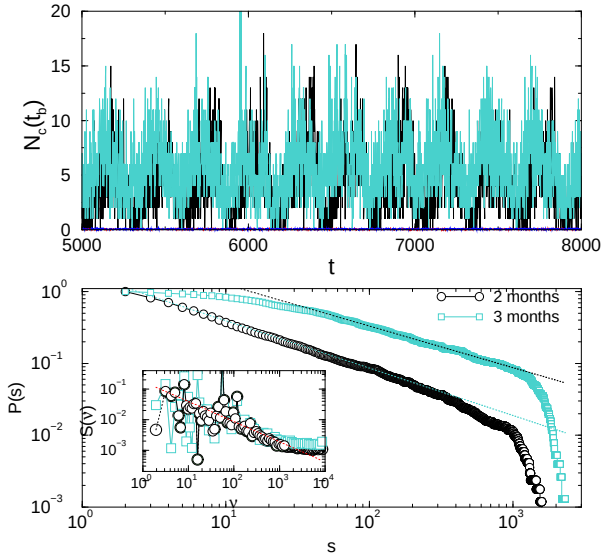


Figure 5.4: Time-series (top) of the dialogs from MySpace networks with two different time depths. Distributions of the avalanche size (bottom) obtained from the time-series, and (inset) power-spectrum of the time-series.

clear power-law behavior before a cut-off, with the exponent close to 0.6, and slightly varying with the time-depth of the network. Similarly, the fractality of the time-series of the dialogs along MySpace networks is reduced and, consequently, the power-spectrum shows weaker correlations and the extended region of white-noise at large frequencies, compared with the above studied time-series data from Diggs and Blogs.

5.3 Summary & Future work

Data, Networks & Emotions analysis performed by WP6 in the first reporting period includes a range of the activities of our group, which (i) enabled us to conduct the meaningful analysis of the data based on theory of complex dynamical systems and complex networks and derive the results anticipated in the Deliverable D6.1; and (ii) makes the foundations (i.e., numerical algorithms, methods and theoretical concepts) for a straightforward continuation in the future work on the Project. In particular, we have developed several methodologies and related numerical algorithms (listed in the attachment) for the collection and the analysis of the particular type of data, i.e., high-resolution data from Blogs and Diggs and networked-type of data with the dialogs from MySpace. In addition, we have developed the theoretical concept and the appropriate model to understand the origin and robustness of the observed collective behavior in Cyberspace, specifically on Blogs and Diggs.

It should be stressed that within our network approach a crucial point is to identify the *User Communities*, which make natural social groups in the Cyberspace, and then apply a detailed analysis of the emotional contents, temporal patterns of user behavior, etc., to understand the collective phenomena *within such connected groups of users*. Our methodology developed within this part of Project research enables a comprehensive analysis of large datasets, as demonstrated above and in the related publications [8, 15, 13], which can be summarized in three major parts as follows:

- *Networks & User Communities.* High-resolution streaming data are mapped onto bipartite networks and User-communities detected on them, and ID's of users within these communities and the lists of their actions over time and related ID's of Posts and Comments selected;
- *Emotions within Selected User Communities.* Emotional contents of the texts posted by the selected individual user, user groups, and dynamical user communities arising over time is analysed and the results associated to the original data as additional property of nodes for further (statistical, temporal) analysis and visualization;
- *Self-Organized Criticality with Emotional Avalanches.* The dynamical evolution with avalanches of comments is traced over bipartite networks, in particular the avalanches of mutually connected emotional comments, as demonstrated in our analysis for the user communities grouped around certain sets of Posts (and individual but very popular posts). The tightly connected user communities and the posts keeping them together are identified as *functional units on the bipartite networks*. This provides a topology evidence in addition to the number of other quantitative evidences that we found for the self-organized criticality with (emotionally charged) avalanches of comments. The set of the *control parameters* that have potentials to tune such self-organized critical states are determined based on the data-analysis and the theoretical model built on the real data and network, are summarized in Table 2. The numerical values of the control parameters are expected to depend on the size of community (including both users and posts) and internal connectivity of the network and potentially on the geographical location of the Blogsite.

In the next period, applying the methodology which we have developed so far, we intend to expand the research in the following directions:

- Complete and analyze the MySpace datasets; Analyze other datasets from Blogs including hyperlinking between different Blogsites, microblogs, etc., to prove the statistics and dynamical stability of the self-organized emotional states in relation to the actual control parameters;
- Developing an agent-based model of blogging and improve the present cellular-automata modeling related to the data by introducing the psychological profiles of the users, according to the measurements of WP7 and the emotional classifier with the extended scale $(-5, +5)$ of emotion, currently developing by WP3. At this level of modeling we expect to better understand the underlying mechanisms [11] of the dynamics and uncover a full range of the emotional activities within the critical avalanches that we are detecting;
- Set up an automated analysis which incorporates all our algorithms along the line discussed above: $Data \Rightarrow Networks\&Communities \Rightarrow Criticality$ and back, in order to improve search in real time for posts, users, and user communities, where relevant events occur, and predict their potential future development.

D6.1: Conclusions

The dynamic critical states with emotional avalanches may occur within user communities arising in the networks of popular posts on Blogs and discussion-driven Diggs. Close-up analysis of the data indicates the self-organized criticality in several measures (avalanche size distributions, $1/\nu$-noise in the fluctuations of the number of emotional comments, Taylor scaling, and time-spacing of the avalanches). These dynamical states depend on several control parameters, which are readily determined from the respective dataset. On the quantitative basis, the critical states are most sensitive to the fraction of very active users, who may give raise of the dissemination parameter over its critical value λ_c for that community (in the case of popular Diggs, estimated λ_c corresponds to approximately 2% of such users), where the supercritical states occur with a number of very large avalanches. These users can be easily identified by their IDs and action times in the data, which gives a possibility to influence their activity and thus the emergent collective behavior. Other modes of the control, e.g., planting positive comments, changing the delay time of user actions, etc., which are captured by other control parameters in our analysis, have much lesser effects on the course of the dynamics. On the qualitative basis, the comments with negative emotions (critique) seem to have much profound effects on the occurrence of the critical states in the considered datasets (different situation is expected in the case of unpopular Posts and in communications within 'friendly' networks like MySpace).

6

List of Algorithms, Attached material and Online-information

6.1 List of numerical algorithms by WP6

(a) Algorithms developed by WP6 (JSI group):

- Python scripts for Blogs data crawling;
- Python script for MySpace networked data crawling;
- Number of C++ codes for:
  - Data mapping onto bipartite graphs; Topology analaysis of the graphs and their projections;
  - Spectral analysis of projected and/or bipartite networks and identification of communities;
  - Statistical analysis of data; Extraction of control parameters from the sets of data; Extraction and analysis of time series for different types of quantities;
  - Simulations of CA-type update rules on real networks and analysis of the simulated data;

(b) Algorithms from other sources used by WP6 (JSI group):

- Emotional classifier with Objective and ± 1Emotional contents (provided by WP3);
- Pajek software for graph visualization (by Pajek); Other graphic packages within Matlab, GNU, xmgr,...;
- SQLite for managing the crawled data from MySpace;
- MEncoder, video decoding, encoding and filtering tool released under the GNU General Public License;

6.2 Movie: evolution of a network with users and emotional comments

Movie describes the evolution of a User community on popular Blogs: Appearance of the Users and Posts and Comments linking the Users are shown in series of steps of the evolution in real-time; Collor of the Comments indicates their emotional content classified as positive (red), negative (black) or neutral (white). Link: http://www-f1.ijs.si/$\sim$tadic/projects/pub/blognet.avi

6.3 List of published science papers, draft, and online-data by WP6

Publications in Journals and Conference presentations related to our results obtained within the CYBEREMOTIONS project (links to the online-data are also available on http://www-f1.ijs.si/∼tadic/projects/cybere.html):

- M. Mitrović , G. Paltoglou, and B. Tadić. Networks and emotions in user communities at popular blogs. DRAFT , 2010.
- M. Mitrović and B. Tadic, Bloggers behavior and emergent communities in Blog space Europ. Phys. J. B: online version: DOI=10.1140/epjb/e2009-00431-9
- M. Mitrović and B. Tadić, Spectral and Dynamical Properties in Classes of Sparse Networks with Mesoscopic Inhomogeneities, Phys. Rev. E vol.80, 026123-1–12 (2009)
- J. Grujić, M. Mitrović and B. Tadić, Mixing patterns and communities on bipartite graphs on web-based social interactions, Proceedings of Digital Signal Processing, 2009 16th International Conference, 5-7 July 2009, Santorini, Greece, ISBN: 978-1-4244-3297-4, DOI: 10.1109/ICDSP.2009.5201238 Current Version Published: 2009-08-18
- B. Tadić, Blogs Dynamics and User Communities , Modelling Science: Understanding, forecasting, and communicating the science system, 6-9 October, Amsterdam, Netherlands, 2009
- M. Mitrović, B. Tadić, Finding Strucure in Blogs: Bipartite Network Analysis, VALUETOOLS 2009 October 20-22, 2009 - Pisa, Italy Fourth International Conference on Performance Evaluation Methodologies and Tools;
- B. Tadić, Modeling Traffic on Networks as Complex Dynamical System, Colloquium, Lakeside Labs, Klagenfurt, 9. November 2009,
- M. Mitrović, Emotions & user communities in Blogs and Diggs : presented at The CyberEmotions Workshop, 21-23 January, Wolverhampton, UK. 2010.
- M. Mitrović, Spectral analysis of networks reveals communities in complex systems data, COST action NP0801 Meeting: Physics of Competition and Conflicts [and] NET 2009: evolution and complexity, Rome, May 28th-30th, 2009.

Bibliography

[1] A. Corral. Long-term clustering, scaling and universality in the temporal occurrence of earthquakes. *Phys. Rev. Lett.*, 92:108501, 2004.

[2] R. Crane, F. Schweitzer, and D. Sornette. New Power Law Signature of Media Exposure in Human Response Waiting Time Distributions. *arXiv:0903.1406*, 2009.

[3] D. Dhar. Self-organized critical state of sandpile automaton models. *Phys. Rev. Lett.*, 64:1613, 1990.

[4] D. Dhar. Theoretical studies of self-organized criticality. *Physica A*, 369:29–70, 2006.

[5] S. N. Dorogovtsev, A. V. Goltsev, J. F. Mendes, and A. N. Samukhin. Spectra of complex networks. *Phys. Rev. E*, 68(4):046109–+, 2003.

[6] Z. Eisler, I. Bartos, and J. Kertész. Fluctuation scaling in complex systems: Taylor's law and beyond. *Advances in Physics*, 57:89–142, 2008.

[7] T. S. Evans and R. Lambiotte. Line Graphs, Link Partitions and Overlapping Communities. *Phys. Rev. E*, 80(1):016105, 2009.

[8] J. Grujic, M. Mitrovic, and B. Tadic. Mixing patterns and communities on bipartite graphs on web-based social interactions. page DOI: 10.1109/ICDSP.2009.5201238, 2009.

[9] J.H. Jensen. *Self-organized Criticality. Emergent Complex Behavior in Physical and Biological Systems*. Cambridge University Press, 1998.

[10] J.H. Jensen. *Self-Organizing Complex Systems: in "Chance Discovery", Ed. Ohsawa*, pp.44-61. Springer, 2003.

[11] J. Kleinberg. The Convergence of Social and technological Networks. *Communications of the ACM*, 51:66, 2008.

[12] R. Lambiotte and M. Ausloos. Uncovering collective listening habits and music genres in bipartite networks. *Phys. Rev. E*, 72:066107, 2005.

[13] M. Mitrović, G. Paltoglou, and B. Tadić. Networks and emotions in user communities at popular blogs. *IJS preprint*, 2010.

[14] M. Mitrović and B. Tadić. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Phys. Rev. E*, 80(2):026123–+, 2009.

[15] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in blog space. *Eur. Phys. Journal B*, 73:293–301, 2010.

[16] T Sadhu and D. Dhar. Steady state of stochastic sandpile models. *J. Stat. Phys.*, 134:427, 2009.

[17] Dj. Spasojević, S. Bukvić, S. Milošević, and H.E. Stanley. Scaling of barkhausen noise in disordered ferromagnets. *Phys. Rev. E*, 54:2531, 1996.

[18] B. Tadić. Nonuniversal scaling behavior of barkhausen noise. *Phys. Rev. Lett.*, 77:3843, 1996.

[19] B. Tadić and D. Dhar. Emergent spatial structures in critical sandpiles. *Phys. Rev. Lett.*, 79:1519, 1997.