UNIVERZITET U BEOGRADU
FIZIČKI FAKULTET

DOKTORSKA DISERTACIJA

STRUKTURA I DINAMIKA TEHNO-SOCIJALNIH MREŽA

Student: Marija Mitrović

Mentor: Prof. dr Bosiljka Tadić

Beograd, 2012

UNIVERSITY OF BELGRADE
FACULTY OF PHYSICS

DOCTORAL THESIS

STRUCTURE AND DYNAMICS OF TECHNO-SOCIAL NETWORKS

Student: Marija Mitrović

Mentor: Prof. dr Bosiljka Tadić

Beograd, 2012

First, I would like to thank my advisor Prof. Dr Bosiljka Tadić for her advice and support. This thesis would not have been possible without her and without freedom and encouragement she has given me from the beginning of our cooperation, four years ago.
The results presented in this thesis are part of my research work on the project CYBEREMOTIONS. During the work on this project I had the opportunity to cooperate with excellent scientists from different fields. Special thanks to Dr Georgios Paltoglou, who provided the part of the data (Digg dataset) and emotional classifier, by which it was possible to obtain the emotions expressed in the communication on the Web.
I would like to thank to all employees at the Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, as well as all other people who work at the institute, for inspiring work atmosphere and help. My first research years I spend in Scientific Computing Laboratory at the Institute of Physics in Belgrade, where I had an opportunity to work and learn from three accomplished scientist Dr Aleksandar Belić, Dr Aleksandar Bogojević and Dr Antun Balaž. I would especially like to thank to my friend Antun Balaž, for all advices and support from the start of my research career.
I owe great gratitude to my family, my parents Slavica and Miodrag and brother Marko, for their unconditional support both in the professional and personal life. Living abroad is not easy but the people around us can make it nicer. I was fortunate to meet great people and colleagues, Marko and Reuben, who made, with their conversations and company, every day at the institute more substantial. Radost, Olivera, Biljana and Rajko showed me how to be more optimistic and how to walk trough life with the smile on my face. Together with Branko, they were like family to me. I owe much gratitude to Branko, for his support and help to discover hidden part of my personality during these two years of our acquaintance and friendship.
In the end, I would like to thank to the Institute Jožef Stefan, for providing me a place to work and all necessary resources and support. The research presented in this thesis would be possible without funding from the European Communitys Seventh Framework Programme FP7-ICT-2008-3 under grant agreement no. 231323 (CYBEREMOTIONS).

Ljubljana, December 2011 *Marija Mitrović*

Prvo bih želela da se zahvalim svojoj mentorki Prof. dr. Bosiljki Tadić za savete i podršku. Ova teza ne bi bila moguća bez nje kao i slobode i ohrabrenja koje mi je pružila od početka naše saradnje, pre četiri godine.
Rezultati koji su predstavljeni u ovoj tezi su deo mog istraživačkog rada na projektu CYBEREMOTIONS. Tokom rada na ovom projektu imala sam prilike da saradjujem sa odličnim naučnicima iz različitih oblasti. Posebno se zahvaljujem dr Jorgosu Paltoglou, koji je obezbedio deo podataka kao i klasifikator emocija, uz pomoć koga smo bili u mogućnosti da dobijemo emocije koje su izražene u komunikaciji na Webu.
Takodje bih želela da se zahvalim svima zaposlenima na Odseku za teorijsku fiziku, Instituta Jožef Stefan u Ljubljani, kao i svim ostalim ljudima zaposlenim na institutu za kreiranje inspirišuće radne atmosfere i pomoć. Svoje prve godine kao istraživač provela sam u Laboratoriji za primenu računara u nauci, Instituta za Fiziku u Beogradu, gde sam imala prilike da saradjujem i ucim od trojice istaknutih naučnika, dr Aleksandrom Belićem, dr Aleksandrom Bogojevićem i dr Antunom Balažom. Posebno se zahvaljujem svom prijatelju Antunu, na svim njegovim savetima i podršci od početka moje istraživačke karijere.
Posebnu zahvalnost dugujem svojoj porodici, mojim roditeljima Slavici i Miodragu i bratu Marku, na bezrezervnoj podršci kako u profesionalnom tako i u ličnom životu. Život u inostranstvu nikada nije lak ali ga ljudi oko nas mogu učiniti lepšim. Imala sam sreće da upoznam dvojicu kolega, Marka i Rubena, koji su kroz razgovore i druženje učinili da svaki dan na institutu bude sadržajniji. Olivera, Radost, Biljana i Rajko su mi pokazali kako da budem optimističnija i kako da kroz život hodam sa osmehom na licu. Zajedno sa Brankom oni su činili moju porodicu ovde u Sloveniji. Branku sam neizmerno zahvalna na njegovoj podršci i pomoći tokom našeg dvogodišnjeg poznastva i prijateljstva, da otkrijem i razvijem jedan skriveni deo moje ličnosti.
Na kraju bih želela da se zahvalim Institutu Jožef Stefan, na obezbedjivanju mesta za rad kao i svih neophodnih resursa. Istraživanje predstavljeno u ovoj tezi ne bi bilo moguće bez finansiranja od strane projekta FP7-ICT-2008-3 grant 231323 (CYBEREMOTIONS) u okviru programa FP7 Evropske Unije.

Ljubljana, December 2011 *Marija Mitrović*

Contents

1 Introduction **11**
- 1.1 Techno-Social interactions . . . 12
  - 1.1.1 On-line communications . . . 12
  - 1.1.2 Role of emotions in on-line communication . . . 13
- 1.2 Complex Networks . . . 14
  - 1.2.1 Community structure and methods . . . 15
- 1.3 Modeling human dynamics . . . 17
  - 1.3.1 Cellular automaton model . . . 18
  - 1.3.2 Agent based modeling . . . 19
- 1.4 The hypothesis . . . 21
- 1.5 The structure of the thesis . . . 22

2 Quantitative methods **23**
- 2.1 Complex networks approach . . . 23
  - 2.1.1 Topological properties . . . 25
  - 2.1.2 Eigenvalue spectral analysis method for community structure 29
- 2.2 Statistical analysis approach . . . 37
- 2.3 Temporal patterns . . . 37
- 2.4 Time-series analysis . . . 39

3 Techno-social networks of Blogs and Digg **43**
- 3.1 Data structure . . . 43
  - 3.1.1 Other on-line social networks . . . 45
  - 3.1.2 Data collection . . . 46
  - 3.1.3 The Emotion Classification . . . 47
- 3.2 Structure of networks . . . 53
  - 3.2.1 Data mapping . . . 53
  - 3.2.2 Topology of networks and their projections . . . 55
  - 3.2.3 Community structure of techno-social networks . . . 59
  - 3.2.4 Emotional driven communities on popular posts . . . 65
- 3.3 Temporal patterns . . . 66

3.3.1 Temporal patterns of User behavior . . . 66
3.3.2 Temporal patterns on Posts . . . 69
3.3.3 Temporal patterns of emotions on popular posts . . . 70
3.4 Evidence of self-organized critical state . . . 73

4 Models of collective behavior at Blogs 77
4.1 Modeling avalanche dynamics . . . 77
4.1.1 Model rules . . . 78
4.1.2 Simulation results: importance of dissemination . . . 81
4.2 Agent-based model on bipartite networks . . . 83
4.2.1 Emotional states of individual agents . . . 85
4.2.2 Model rules . . . 88
4.2.3 Model parameters . . . 90
4.2.4 Simulated temporal patterns . . . 93
4.2.5 Simulation of emergent bipartite networks and communities of the emotional agents . . . 96
4.2.6 Temporal patterns of emotional communities . . . 103
4.2.7 Circumplex map . . . 105

5 Summary and Conclusions 109
5.1 Network properties and emotions . . . 110
5.2 Self-organized criticality of emotional behavior . . . 111
5.3 Future work . . . 112

A Eigenvalue spectral analysis method 115

B Data collection 121

Abstract

On-line communications at Web portals represent technology-mediated user interactions, leading to massive data and potentially new techno-social phenomena not seen in real social mixing. Large-data resulting from these techno-social interactions provide the ultimate source of information to study emergent social behavior. The interactions of users via posts are indirect, suggesting the importance of the contents of the posted material, such as subject or emotion. In this thesis we focus on the role of emotions in dynamics of on-line social systems. We present a systematic way to study empirical data from Blogs and Digg, which combines the approaches of *theory of complex networks* and *physics of complex systems*.

The data from Blogs and Digg are mapped onto a bipartite network where users and posts with comments are two natural partitions. The topological properties of obtained bipartite networks and their projections are closely related to dynamical characteristics of considered systems. Specifically, we detect clusters of users, communities, in weighted projections of bipartite networks using spectral analysis method. By analyzing the content of posts (and comments), around which are clustered certain groups of users, we find that there are two different types of communities. For users in communities related to normal posts the subject of these posts is is of great importance, while our analysis of subgraphs on popular posts suggest that emotions expressed in post and comments have important role in evolution of these communities.

Human activity on the Web can be considered as a time series of number of comments. These time signals are analyzed using tools of physics of complex systems. In time series obtained from Blogs and Digg data we observe avalanches of emotional comments exhibiting significant self-organized critical behavior and temporal correlations. We design a data-driven network-automaton model in order to explore the robustness of these critical states. The model is implemented on realistic network which can be, together with oder model parameters, inferred from the empirical data. The simulations results show that dissemination of emotions by a small fraction of very active users appears to critically tune the collective states.

In order to study mechanisms underlying the collective emotional behavior of Bloggers, we design the agent based model with values of parameters obtained from empirical data. The state of emotional agent is quantified with two variables, arousal

and valence, fluctuates in time due to events on posts connected to users, and in the moment of agent's action it is transferred to a selected post. These agents, together with posts, form the bipartite weighted network, which evolves in time due to their actions. We investigate the dependence dynamical and structural properties of system on values of certain parameters. Simulations show that the collective behavior, which we recognize by emergence of communities on the network and the fractal time-series of their emotional comments, is powered by negative emotions (critique).

Apstrakt

Onlajn komunikacija na Web portalima predstavlja tehnološki posredovanu interakciju izmedju korisnika, koja za posledicu ima razvoj novih tehno-socijalnih fenomena koji nisu prisutni u standardnoj socijalnoj komunikaciji. Podaci koji nastaju kao posledica ovih tehno-socijalnih interakcija, predstavljaju jedinstveni izvor informacija za proučavanje pojavnog socijalnog ponašanja. Interakcije korisnika preko priča (postova, članaka) su indirektne, što ukazuje na značaj njihovog sadržaja, kao što su tema ili emocije. U ovoj tezi je poseban fokus na ulogu emocija u dinamici onlajn socijalnih sistema. Predstavljamo sistematičan način za proučavanje empirijskih podataka sa Blogova i Digova, koji kombinuje metode *teorije socijalnih mreža* i *fizike kompleksnih sistema*.

Podaci sa Blogova i Diga su mapirani na bipartitne mreze, gde korisnici i price zajedno sa komentarima na njima, predstavljaju dve prirodne particije. Topološke osobine dobijenih bipartitnih mreža i njihovih projekcija su usko povezane dinamičkim karakteristikama posmatranih sistema. Konkretno, koristeći metod sprektralne analize, u otežinjenim projekcijama bipartitnih mreža nalazimo grupe korisnika, takozvane zajednice. Analizirajući sadržaj priča (kao i komentara na njima), oko kojih su ovi korisnici klasterisani, nalazimo da postoje dva raličita tipa zajednica. U zajednicama korisnika koji komentarišu normalne postove, tema ovih postova je od velike važnosti, dok analiza podgrafova na popularnim pričama ukazuje da emocije iskazane u njima i njihovim komentarima imaju važnu ulogu u vremenskoj evoluciji ovih zajednica.

Aktivnost ljudi na Web-u može biti pretočena u vremenski signal broja komentara. Ovi vremenski signali mogu biti analizirani metodama statističke fizike kompleksnih sistema. U vremenskim serijama Blogova i Digova opažamo da lavine emotivnih komentara ispoljavaju značajno samo-organizovano kritično ponašanje kao i vremenske korelacije. Da bi smo istražili robustnost ovih kritičnih stanja, razvili smo model baziran na principu ćelijskih automata. Model je implementiran na realnoj mreži koja je, zajedno sa ostalim parametrima dobijena iz empirijskih podataka. Rezultati simulacije pokazuju da diseminacija emocija od strane malog broja veoma aktivnih korisnika utiče na razvoj kolektivnih stanja.

U nameri da što bolje proučimo mehanizme koji vode do kolektivnog emotivnog ponašanja blogera, dizajnirali smo model agenata za koji su vrednosti parametara

takdoje dobijene na osnovu analize empirijskih podataka. Stanja emotivnih agenata su kvantifikovana dvema varijablama, *uzbudjenjem* i *valencom*, čije fluktuacije u vremenu direktno zavise od dogadjaja na postovima vezanim za korisnike. Emotivno stanje se u trenutku korisnikove aktivnosti prenosi na post koji on komentariše ili piše. Agenti, zajedno sa postovima, sačinjavaju bipartitnu otežinjenu mrežu, koja evoluira u vremenu zahvaljujući njihovoj aktivnosti. U ovoj tezi, proučavamo zavisnost dinamičkih i strukturnih osobina sistema od vrednosti odredjenih parametara. Simulacije pokazuju da kolektivno ponašanje, koje se reflektuje kroz zajednice korisnika i fraktalnu strukturu vremenskih serija emotivnih komentara, je vodjeno negativnim emocijama (kritikama) ispoljenim u komentarima.

Chapter 1

Introduction

The Internet experience in recent years has revolutionized the mechanisms that an individual can exploit to participate in global social dynamics. New techno-social phenomena which emerge on the web [1, 2, 3, 4] boost an intensive multidisciplinary research with interconnected contributions from the physics of complex dynamical systems, computer science, and social science. For example, in technology research new generations of of services are developing in the human capabilities in a service-oriented manner [5]. Algorithms for safe and efficient information collection and processing are developed. Activity of users in virtual life is influenced by real-time events. Recent events such as the revolution in North African countries show that human activities in the virtual society also have an impact on every-day life. For this reasons studying of the Web has become concern of social sciences and every day's practice. The research of the Web concerns with understanding its structure [6] and the underlying evolution mechanisms [7] as well as the emergent social phenomena among Web users [1, 3]. For this reasons systems formed by on-line communication are studied using methods and tools of complex dynamical systems.

The internal structure of a complex system manifests itself in correlations among its constituents. In complex systems interactions between constituents cause *collective modes* having special statistical properties which reflect underlying dynamics. To quantitatively describe such collective behavior in human systems, using methods of statistical mechanics and network theory, one needs to analyze large datasets. The data collected from the Web provide the basis to study human behavior at an unprecedented level of details. For instance, from the high-resolution data stored at various Web portals (social networks, Blogs, forums, chat-rooms, computer games, etc) information related to user preferences, patterns of behavior, attitudes, and emotions can be inferred for each individual user and user communities gathered around certain popular subjects [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Analysis of the data collected from Web portals enables us to study communication among

humans but also a new phenomena which can arise as a consequence of the users interactions on contemporary communication networks and social media.

1.1 Techno-Social interactions

Communication is the activity of conveying meaningful information. It requires a sender, a message, and an intended recipient, although the receiver need not be present or aware of the sender's intent to communicate at the time of communication; it can occur across vast distances in time and space, and requires that the involved parties share an area of communicative commonality. The communication can be verbal and non-verbal and we use it to share information, experience, opinions, emotions, etc. It can be synchronized or asynchronized, face-to-face (F2F) or computer-mediated (CM). By communicating with each other human are getting involved in different relationships, forming groups (communities) characteristic for collective dynamics.

1.1.1 On-line communications

Computers and electronic networks have changed the way we are communicating with each other. Without having to bother with stamps, envelopes, and the delay in postal mail (called "snail mail" by electronic mail enthusiasts), millions of people interact via e-mail on every-day basis. The Web 2.0 applications represent a set of tools that enable the masses to easily create content on the WWW, in the form of Blogs, social networks, video and photo collections, and simple application creation frameworks. Their development has amplified the importance of relationships between users that are represented in social networks. During this activities, users create a huge amount of data which can be used for the analysis human dynamics on the Web. The emergence of socially rich computing applications with millions of users allows us to ask questions that were impossible to answer before as large scale human social dynamics data was practically impossible to collect.
An important feature of the on-line communications is that user interactions are mediated by the posted material, e.g., the text of posts and comments on the Blogs and Digg, studied here. The computer mediate interactions are characterized with the *absence of person*. Such communication is usually asynchronous and is devoid of gestures, touch, body language or facial expression. Despite that, the posted text may in different ways affect the actions of the users who read it, depending on the information that the text contains, but also by featuring certain aesthetic, moral or emotional contents [20, 21]. All this can lead to the collective behavior which have not been yet understood. Therefore, scientific data analysis and modeling from the

point of view of the complex evolving systems appears as a necessary step towards better understanding of the social phenomena in cyberspace.

1.1.2 Role of emotions in on-line communication

Human emotions and contacts

Study of emotions started with Darvin in 19th century who setup biological framework. Based on this, psychologists have researched affective processes with regard to (i) causes, (ii) mental processes and bodily systems involved, (iii) intra- and interpersonal regulation, (iv) communication, and (v) consequences. The majority of theories and research during 20th century have focused on psychological processes within the individual, neglecting the complex behavior that emerges when individuals interact. Human emotions are strongly influenced by social contacts. The studies on *emotional contagion* show that different type of moods, positive and negative, are transmitted on short timescales between people in different types of close contacts [22, 23, 24, 25]. The way we manifest our emotions and how we perceive the ones expressed by others leads to complex social behavior. This behavior is poorly understood, mainly due to the lack of empirical data.

Emotions in on-line communication

Our ability to express and accurately assess emotional states is central to human life. Emotions are typically expressed trough a variety of non-linguistic mechanisms: laughing, smiling, vocal intonation and facial expression. This relationship between emotions and nonverbal communication has been examined by various research [26]. On the other hand how emotions are reflected verbally is not so well understood [27]. The fact that computer mediated communication is devoid of audio and visual signs, has frequently led to the assumption that text-based communication has a reduced capacity for emotional exchange. Research in the social information processing theory [28] suggests that users can employ words and emoticons, which are characteristic for computer-mediated communication, to convey relational information that may normally be transmitted via nonverbal cues in face-to-face contexts. For example, the expression of affinity or dis-affinity towards a partner can be equally expressed in CM or F2F communication [29]. Ability of users to effectively express their liking or disliking of one another in mediated environments indicates that users can also reveal their current emotional state, such as positive and negative valence emotions, in CMC communication. Recent studies show that the emotions expressed

in the text (or other posted materials) play an important role in the on-line 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 [30] and observed on the level of large-scale social effects [21, 31].
The studies of affective interactions on the Web are essential for understanding of role of emotions in social dynamics [32, 12, 16, 17, 18, 13]. These studies are based on the analysis of large data sets which in addition to various information about user activity contain emotional content of textual messages. The necessity to extract emotional content from the text has led to the development of a number of algorithms for detection of positive and negative sentiment. On the other hand, this development has made large-scale on-line research possible, such as predicting elections by analyzing sentiment in Twitter [33], and diagnosing trends for happiness in society via blogs [10] and Facebook status updates[34].

1.2 Complex Networks

Human social interaction can be studied trough social networks, with humans as nodes and their relationships represented by links. Social networks, i.e., their structure and influence on social dynamics, are subject of investigations in diverse fields, including sociology, economics, physics, mathematics and public health.
The networks are all around us. As individuals, we are members of network of social relationships of different kinds, as biological systems we are result of a network of biochemical reactions. Any system composed of interacting elements can be represented by complex networks [35]. These networks can be of different nature. Electric power grids, the Internet, highways or subway systems, and neural networks are tangible objects in the Euclidean space. On the other hand, the networks of acquaintances or collaborations between individuals or networks of biochemical reactions, are entities defined in an abstract space.
For a long time the study of networks has been mainly the domain of a branch of discrete mathematics known as *graph theory*. Since its beginning graph theory has provided answers to many practical questions such as what is the maximum flow per unit time from source to sink in a network of pipes, how to color the regions of a map using the minimum number of colors so that neighboring regions receive different colors, or how to fill n jobs by m people with maximum total utility. Parallel with the theory of graphs, the study of networks has been developed in social sciences. Social networks analysis focuses on relationships among social entities, as communication between members of a group, trades among nations, or economic transactions between corporations. In the last two decades the interest and research in this field moved toward networks whose structure is irregular, complex and dy-

namically evolving in time. The development of computer facilities and power has also directed the research toward systems with thousands or millions of nodes such as transportation networks, phone call networks, the Internet and the World Wide Web, the actors collaboration network in movie databases, scientific co-authorship and citation networks from the Science Citation Index, neural, genetic, metabolic and protein networks.
The research in field of complex networks started with an effort to define a new concepts and measures to characterize the topology of real networks. The analysis of different real-world networks has resulted in the identification of a series of unifying principles and statistical properties common to most of them. It was found that certain topological measures that allow us to study structural properties of networks exhibit similar behavior independent of systems nature. For instance, a degree distribution ($P(q)$), where degree of the node is the number of its directed connections to other nodes, significantly deviates from the Poisson distribution expected for a random graph. In many cases, it exhibits a power law (scale-free) tail with an exponent between 2 and 3. Beside this, real world networks are characterized by correlations in the node degrees [36], by short average distance between any two nodes [37], (known as *small-world property*), and presence of a large number of short cycles, *specific motifs* [38, 39], and *communities* [40]. All these measures allow us to quantitatively describe and better understand dynamics of complex systems.

1.2.1 Community structure and methods

The structure of the real network is strongly connected with the function and dynamic of complex system [35, 41]. For example the communities or subgraphs play an important role in the networks complexity along the line from the local interactions to emergent global behavior, both in the structure and the function of networks [35, 42]. Understanding the mesoscopic structure of networks in both topological and dynamical sense is, therefore, of paramount importance in the quantitative study of complex dynamical systems. For this reasons a community detections in networks is one of the important steps in understanding complex dynamical systems represented by them. Unfortunately, the problem of graph clustering, is actually not well defined [40]. Specifically, there is no rigorous definition of community and partition. Different concepts of communities result in different recipes and algorithms for their detection [40]. The most common definition of communities is based on edge density, inside versus outside the community. The same definition is used in this thesis for description of user activity on the Web [12, 16, 17, 18]. One can also define a community as group of similar nodes according to some additional property. This definition is used in methods for data clustering [43].
Techniques for detection of communities in networks, which incorporate the concept

of edge density, can be categorized into three different categories, *hierarchical*, *partitional* and *spectral* clustering [40]. Selection of a particular method for subgraph detection in a particular network depends on its topological properties, such as edge density and/or its size.

Most of the methods use the centrality measures, topological [44] or dynamical [45] flow, and are based on the maximal-flow-minimal-cut theorem [46]. The theorem states that the minimum cut between any two vertices of a graph, i.e., any minimal subset of edges whose deletion would topologically separate these two vertices, carries the maximum flow that can be transported from one to another node across the graph. This theorem has been used to determine minimal cuts from maximal flows in clustering algorithms. There are several efficient routines [47, 48] that use maximum flows to identify communities in the graph. The communities in this algorithms are defined to be *strong*, i.e. the internal degree of each vertex must not be smaller than its external degree [49]. An artificial sink t is added to the graph and one calculates the maximum flows from a source vertex s to the sink t: the corresponding minimum cut identifies the community of vertex s, provided s shares a sufficiently large number of edges with the other vertices of its community, otherwise one could get trivial separations and meaningless clusters.

Another effective approach for graph partitioning is based on statistical methods of maximum-likelihood [50, 51]. These methods use machinery of mixture models and expectation-maximization algorithm for finding the best partition of the graph. The nodes are grouped according to their patterns of connections, i.e. nodes with similar sets of first neighbors according to their strongest connections are assigned to the same community.

Several methods are considered as dynamical algorithms and employ the processes running on the graph. For example, random walkers on complex networks spend more time within subgraph and move more rarely from one subgraph to another [40]. By analyzing the random walk dynamics one can deduce which nodes belong to the same community. Synchronization [52] is an emergent phenomenon occurring in systems of interacting units and is ubiquitous in nature, society and technology. In a synchronized state, the units of the system are in the same or similar state(s). Synchronization has also been applied to find communities in graphs. If oscillators are placed at the vertices, with initial random phases, and have nearest-neighbor interactions, oscillators in the same community synchronize first, whereas a full synchronization requires a longer time. So, if one follows the time evolution of the process, states with synchronized clusters of vertices can be quite stable and long-lived, which enables the extraction of communities [53, 54].

Synchronization and random walk dynamics are related to spectral clustering. In fact, normalized Laplacian, used in eigenvalue spectral analysis method, is related to random walk dynamics [55], while the eigenvalues of Laplacian matrix, $L_{ij} = q_i - A_{ij}$, are inversely proportional to synchronization time for a certain group of nodes in the

network. Spectral properties of Laplacian matrix are closely related to structural properties of the networks and can be used for identification of network community structure [55]. Specifically, the eigenvectors of smallest non-zero eigenvalues are localized on network subgraphs [55]. The detailed description of this method is given in Section 2.1.2.

1.3 Modeling human dynamics

The study of complex systems in a unified framework has become recognized in recent years as a new scientific discipline, the ultimate of interdisciplinary fields. It is strongly rooted in the advances that have been made in diverse fields ranging from physics to sociology, from which it draws inspiration and to which it is relevant.
Many of the systems that surround us are complex. Understanding their properties is a central motive for all involved scientific disciplines. The goal of understanding their properties motivates much if not all of scientific inquiry. Despite the great complexity and variety of systems, universal laws and phenomena are essential to our inquiry and to our understanding. All complex systems exhibit some common characteristics:

- They consist of large number of interacting elements - *agents*.
- They exhibit *emergence*, a self-organized collective behavior, which is difficult to anticipate from the knowledge of actions of isolated agents.
- Their emergent behavior does not result from the existence of central controller.

The appearance of emergent properties is the most distinguishing feature of complex systems. In these systems, even perfect knowledge and understanding do not give predictive information [56]. For this type of systems computational models represent the optimal means of prediction.
A model is a simplified mathematical representation of the system, in which only relevant features that are thought to play an essential role in the interpretation of the observed phenomena could be retained. A simple model, if it captures the key elements of a complex system, may elicit highly relevant questions.
Depending on the system features one can use either *analytical* or *computational* models to study its dynamics. Analytical models are used when the solution of set of equations which describe changes in the system can be expressed as a mathematical analytic function. Sometimes, even simple systems with a small number

of parameters and interactions, can exhibit non-linear features. The computational models can be used for illustration and tracking of these features. In computational models, the dynamic of every single object in the system is determined by set of rules. Numerical simulations enables us to follow the activity of each component in time and assist us in understanding emergent phenomena.
In this thesis we use two different concepts for modeling emotional collective behavior:

- *cellular automaton (CA)* - is a discrete dynamical system composed of identical cells (points in regular spatial lattice) which can have a finite number of states. The state of each cell at given time depends only on its own state and the states of its nearby neighbors at previous time step. The state of the whole system evolves in discrete time steps [57].

- *agent-based model (ABM)* - is microscopic model that describes a system from the perspective of its constituent units - agents. It consists of a system of agents and the relationships between them. In some ABM, agents are capable of evolving, which enables the emergence of unanticipated behavior.

Both types of models are capable of spontaneous developing of dynamical *self-similar* patterns. *Self-similarity* and *self-organization* are well characterized collective phenomena.

1.3.1 Cellular automaton model

Social, biological and many other systems in nature exhibit extremely complex behavior generated by cooperative effects. The cellular automata are discrete dynamical systems composed of many identical interacting components capable of complex behavior [57]. Cellular automata consists of certain number of elements, called cells if it is defined on a regular grid, who can adopt a finite number of states. The state of the cell depends on the state of the neighboring elements based on set of fixed rules.
The basic idea of cellular automata models can be explained on the most famous one "The game of life" constructed by John Conway [58] which is developed for studying population dynamics. The game is defined on two dimensional orthogonal grid of square cells. Each cell can adopt one of two possible states, *live* or *dead*, and interacts with its eight neighbors, i.e. horizontal, vertical and diagonal. Weather the cell will be live or dead depends on the state of its neighbors. The live cell with less than to live neighbors dies, the one with two or three — survives to the next generation, while the cell with more than three neighboring cells — dies as a consequence of overcrowding. A dead cell becomes alive if and only if three of its neighbors are alive.

Despite their apparently simple definition, based on local rules, cellular automata can show very complex dynamical behaviors, even in the case of the most simple $1D$ cellular automata with two neighbors and two states. The CAs are applied for modeling and understanding of different dynamical systems. In the creation of CAs models one can distinguish two main approaches: forward and backward. The forward, theoretical, approach concerns the study of transition rules of a given cellular space in order to establish its intrinsic properties such as dynamical behavior or pattern formation. The backward, practical, approach regards the design of transition rules sets of a designed cellular space in order to match the *right* behavior of the CA system of a given complex system (physical, biological, social and so on). These two approaches are interconnected, i.e., sometimes actual applications of CA are directly derived from the theoretical results. In other cases, set of backward rules successfully mimics the dynamics of a real complex system and theoretical rules of that activity have to be investigated.
The cellular automata have been developed and used in many different fields, such as statistical physics [59] or biology and ecology [60]. They were used for simulation of conceptually different complex systems such as lattice gas model [61], lattice Boltzmann automata for simulation of fluid dynamics [62] or percolation of water [63]. As dynamical systems which are capable of displaying complex behavior, CAs can also provide the insight into self-organized criticality (SOC). Self-organized criticality concerns a class of dynamical systems which naturally drive themselves to a state where interesting physics occurs over a wide range of length and times scales [64, 65]. The SOC systems commonly modeled by CAs are sand-piles [66, 67]. In socio-physics they are used for modeling pedestrian dynamics [68], opinion formation [69] or trading markets [70].

1.3.2 Agent based modeling

Complex system consists of large number of interacting elements capable of self-organized collective behavior [65]. The agent-based model consists of set of interacting agents that encapsulate the activity of various elements that make up the system, emulating the dynamics of the system [71]. It follows from this that the agent-based models (ABM) are suitable for modeling complex systems. In ABM, a system is modeled as a collection of autonomous decision-making entities called agents. Each agent individually assesses its situation and makes decisions on the basis of a set of rules. Agents may execute various actions, such as producing, consuming, selling, expressing the opinion or emotion, or sharing the information. The selection of agent's properties and their activities depend on features of the modeled system. An agent-based model consists of a system of agents and the relationships between them. These interactions between agents are a feature of agent-based mod-

eling, which relies on the power of computers to explore dynamics out of the reach of pure mathematical methods [72, 73]. Even a simple agent-based model can exhibit complex behavior [74] and provide valuable information about the dynamics of the real-world system that it emulates. In addition, agents may be capable of evolving, allowing unanticipated behaviors to emerge. Sophisticated ABM sometimes incorporates neural networks [71, 75, 76], evolutionary algorithms [77], or other learning techniques to allow realistic learning and adaptation.

The main features of AMB model which make them suitable for modeling of complex systems are: (i) they capture the emergent phenomena; (ii) they provide a natural description of a systems; (iii) they are flexible. As was mentioned, emergent phenomena result from interactions of individual entities. An *emergent phenomenon* is a large scale, group behavior of a system, which doesnt seem to have any clear explanation in terms of the systems constituent parts. By definition, the emergent phenomena can not be reduced to the systems parts: the whole is more than the sum of its parts because of the interactions between the parts. It is often counterintuitive. For example, a traffic jam, which results from the behavior of and interactions between individual vehicle drivers, may be moving in the direction opposite that of the cars that cause it. In the ABM one simulates the activity of the system's elements and their interactions, capturing the emergent phenomena. The structure of ABMs makes them the most natural choice for describing and simulating a system composed of active entities, such as traffic jam [78], the stock market [79] or voters [73]. In ABMs it is easy to add more agents or to change levels of description and aggregation, which makes them flexible.

The development of multi agent systems represents a breakthrough in computational modeling in the social sciences [71]. The modeling trough agents allows scientist to extract specific properties of humans, such as emotions, will or subject preference and study individual influence of these properties in society.

The contribution of ABMs to social sciences for a long time was theoretical and abstract. They were used to show how simple rules of interaction could explain macro-level phenomena. Although most of the models were inspired with real social and biological systems, many of them have not been tested and compared with empirical data. For a long time they have been used only as a *proof of concept*, not for prediction or quantitative study of system dynamics. The development of ABMs showed that they are valid technical methodology. This and the availability of large volumes of high-quality data led to development of ABM which were motivated by empirical observations and measurements. The empirical data can be used, both qualitatively and quantitatively, i.e., they can be used for deducing values of input parameters for the model, or as means to test a model, or both.

1.4 The hypothesis

The thesis addresses a number of important questions regarding the properties and patterns of collective behavior of humans on the Web. Specifically, we are interested in finding suitable quantitative measures trough which this collective behavior can be identified. Also, we are interested in the role of emotions in dynamics of techno-social systems.
The dissertation focuses on the methods for quantitative study of dynamics and structure of techno-social interactions with particular emphasis on role of the emotions. We would like to show that *theory of complex networks* and *statistical physics of complex systems* provide appropriate tools for quantitative analysis of collective behavior on the Web. We argue that the data collected from Blogs and similar Web portals can be reliably represented with directed bipartite networks, and that structural properties of these networks and their projections are closely associated with the users activity on the Web. The main activity of humans on the Blogs is comment posting which can be transformed into temporal signal. Our hypothesis is that these signals can be studied and analyzed using the same methods as for earthquake and sand-pile dynamics, or Barkhausen noise. Using the tools of *physics of complex systems* such as time series analysis we would like to investigate dynamical properties of on-line social systems, and how emotions expressed in messages influence dynamical patterns.
We postulate that techo-social systems exhibit *self-organized collective behavior*. *Self-organization* is a process whereby the pattern at the global level of the system emerges solely from interactions from its lower-level components. In self-organizing systems, pattern and organization develop without the intervention of external influences, such as *leader* who directs or oversees the process. *Collective behavior* in social system represents relatively spontaneous and unstructured ways of thinking, feeling, and acting that develop within a group as a result of interaction among participants. This behavior not governed by traditional, established norms and rules. The techno-social systems do not differ in this sense from other complex systems. In this thesis we would like to show, that self-organization and collective behavior powered by emotions, can be identified trough topological properties of directed bipartite networks, such as degree distribution or *community* structure, and trough statistical features of different time signals of the system.
The analysis of empirical data provides information on the structure and dynamics of on-line human interactions in social media. The empirical observations can be used for construction of numerical models, such as cellular automata or agent-based models. These models can be used for testing of some hypothesis and/or for prediction. Our intentions in these thesis is to develop models which can hep us to investigate the role of emotions in techno-social dynamics. We assume that some features of human activity on Blogs and Digg are mostly a consequence of emotional

interaction and that models of systems consisting of emotional agents can exhibit similar behavior which was observed in real systems.
Hear we want to highlight the fact, that suggested methods for quantitative analysis of the empirical data and data driven models, together make a complete methodology for studying the complex human behavior on the Web.

1.5 The structure of the thesis

Our intension in this thesis is to present a comprehensive methodology for studying collective emotions on Blogs and other Web sites where user's interactions are not direct but mediated trough posted material. In Chapter 2 we present some of the methods which are traditionally used for quantitative analysis of complex systems. As it was already stressed, the methodology consists of two major parts. In Section 2.1 we explain the basic concepts of *theory of complex networks* such as: data mapping, important topological properties (Section 2.1.1) and eigenvalue spectral analysis method which will be used for identification of community structure in networks (2.1.2). Important tools which we adopted from *statistical physics of complex systems* are presented in Section 2.2. In Chapter 3 we explain how the data from Blogs and Digg were collected (Section 3.1.2) and how the emotions were classified (Section 3.1.3), so as their description (Section 3.1). The quantitative analysis of these data using the methodology proposed in Chapter 2 are presented in Sections 3.2, 3.3 and 3.4. Specifically, the topological properties of directed bipartite networks and their projections are presented in Section 3.2, while Section 3.3 is about the observed temporal patterns of users behavior and community evolution. The evidence of emotional self-organized critical state of these systems is given in Section 3.4.
In order to study the role of emotions in collective behavior observed in empirical data, we develop two conceptually different models. The descriptions of these models are given in Chapter 4 together with the analysis of simulation results. The network automaton model was developed in order to test how self-organized critical state is developed in the system and how it depends on some system properties (Section 4.1). The agent-based, model of human emotional dynamics which enables more detailed studies is presented in Section 4.2. The summary the main results and conclusion of these thesis so as proposal of future work is given in Chapter 5.

Chapter 2

Quantitative Methods for Study of Human Dynamics on the Web

What is common to ant colonies, human economic and social groups, climate, nervous systems, cells and living things? They all belong to a class of dynamical non-linear systems which exhibit complex behavior. These systems are variously described as being complex, because they have numerous internal elements, and dynamic, because their global behavior is governed by local interactions between the elements, Understanding of complex systems is essential for future technological and scientific development. Quantitative analysis of complex systems is crucial for understanding and description of their dynamics.

In this chapter we will present the methods for quantitative analysis of on-line social behavior, based on *complex network theory* and *statistical analysis*. We will describe some basic concepts for identification and description of human dynamics on the Web.

2.1 Complex networks approach

Many systems in nature and technology can be represented by a large number of highly interconnected dynamical units [35]. The approach to capture global properties of these systems is to represent them as graphs whose nodes represent dynamical units and the links stand for the interactions between these units. Depending on the type of nodes and interactions, one can use different network representation of some system. For example, the system in which the only question is weather two elements interact, is represented with *binary* network. In the cases where interactions are of different strength, the representing network is weighted and the weights of links are proportional to the strength. If the relationships are not symmetrical, the links of

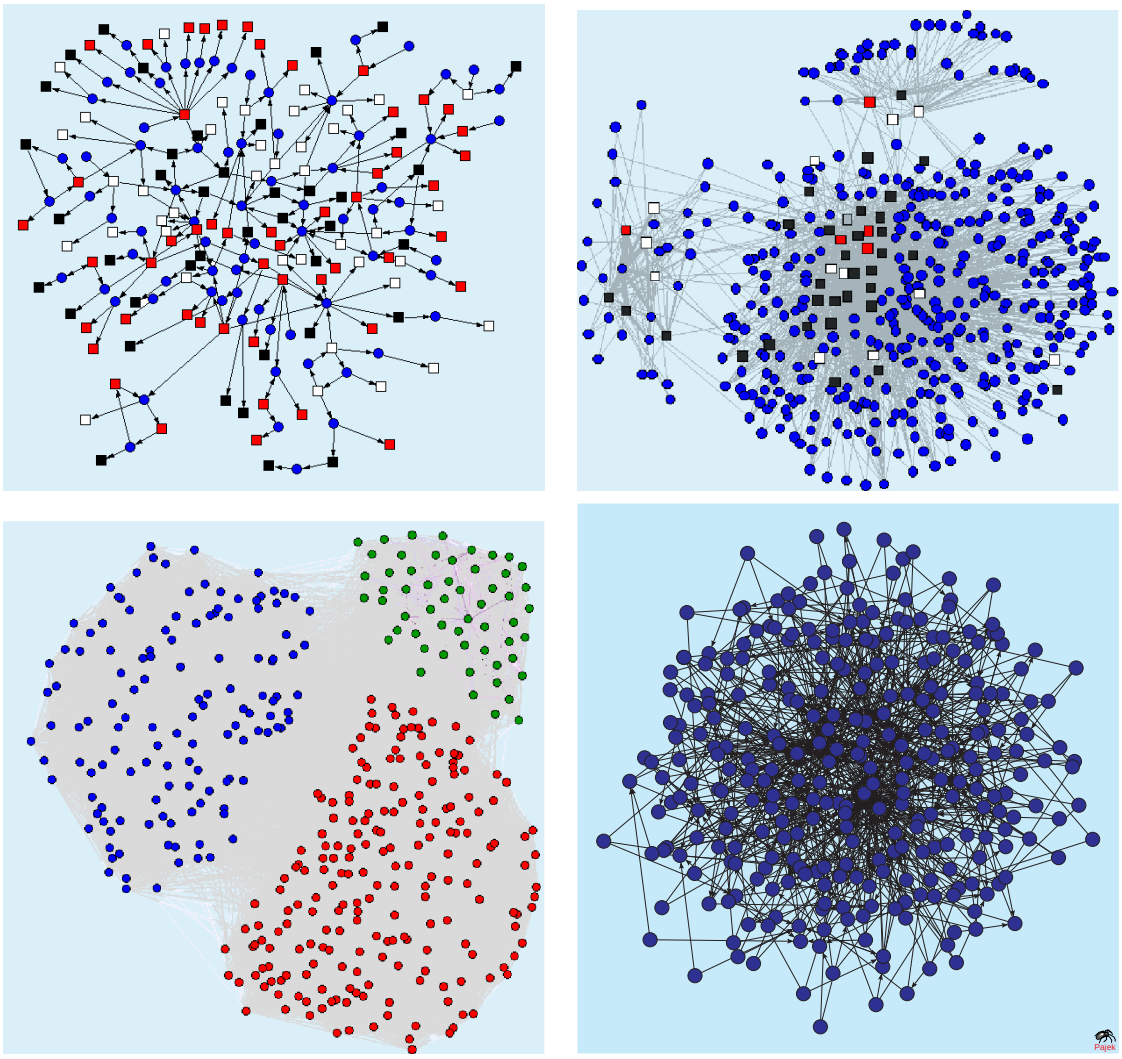

Figure 2.1: (top left) Directed bipartite network related to one Digg story with comments, shown as colored boxes, and users, shown as bullets. (top right) Weighted bipartite networks of popular BBC Blogs with posts as colored boxes, and users as bullets. Weights of the links indicate multiple comments by the user to the post, (bottom left) Monopartite weighted network of users from popular Digg stories. (bottom right) Example of undirected binary network of size $N = 1000$ generated using model of modular networks for parameters $M = 3$ $\alpha = 0.95$ and $P_0 = 0$.

the network have direction and the network is called *directed*, otherwise the network is denoted as *undirected*. Most of the systems in the nature consist of interacting elements of the similar type. For these systems the *monopartite* networks are suit-

able. On the other hand, some of the systems consists of two basic types of elements, which interact only with one of the different nature. The *bipartite* networks are used for representing of this type of systems. Some examples of representing of on-line social systems are show in Fig. 2.1.

The complex systems can consist of millions of elements and are represented with the network of corresponding size. The development of powerful and reliable data analysis tools and computer facilities, have lead to better machinery to explore the topological properties of these networks.

2.1.1 Topological properties

Mathematical representation of networks

Networks are naturally represented in matrix form. A graph of N nodes is described by matrix of size $N \times N$ whose non-zero elements indicate connections between nodes. Depending on type of network we can consider different kinds of matrices:

- *Adjacency matrix*, A, with elements $(0, 1)$, which is suitable for representing networks with links of equal strength. If element $A_{ij} = 1$ there is a link from node i to node j. In the case of undirected networks A is a symmetrical matrix.

- *Matrix of weights*, W, for representation of weighted network. The element W_{ij} denotes the strength of the connection between nodes i and j. Again symmetrical form of matrix W is a consequence of undirected nature of interactions in the system.

Any structural characteristic can be expressed in terms of these matrices. For example a degree of the node, defined in next paragraph, is calculated as $q_i = \sum_j A_{ij}$ or $q_i = \sum_j \theta(W_{ij})$ in the case of weighted network. Here $\theta(x)$ denotes Heaviside step function.

Degree distribution

Degree q_i of the node i is defined as number of first neighbors of that node [35, 80]. The degree distribution $P(q)$ is the probability that randomly chosen node in a network has degree q:

$$P(q) = \frac{< N(q) >}{N} \ . \tag{2.1}$$

Here $< N(q) >$ is the average number of nodes with degree q, averaged over the entire statistical ensemble, with assumption that total number of nodes in each member of the ensemble is the same N. In the case of empirical data, where there is only single realization of the network, graph g, the degree distribution represents the frequency of occurrence of nodes with degree q, $P(g) = \frac{N_g(q)}{N}$ where $N_g(q)$ is number of nodes with degree q. Note that $P_g(q)$ approaches $P(q)$ in the limit of infinite network.

Calculation of the degree distribution is the simplest statistical characteristic. Different nodes in the network have different roles and functions, which is reflected in the value of their degree. Both low- and high-degree parts of distribution are equally important for understanding and description system. For example, classical random graphs have degree distribution which decays rapidly, $P(q) \sim \frac{1}{q!}$, for large q. All their moments $\sum_q q^n P(q)$ are finite even as the network size approaches infinity. The average degree $< q >= \sum_n qP(q)$ represent a typical scale for degree in this type of networks. Research of different real world networks show that hubs occur with noticeable probability and have an important role in network structure [35]. Occurrence of hubs is related with slowly decaying degree distribution, which usually has power-law asymptotic behavior, $P(q) \sim q^{-\gamma}$, in the limit of the infinite network. Higher moments of this distribution diverge with network size, meaning that there is no typical node degree in the network. For this reasons, these networks are called *scale-free* networks.

In finite-size networks, degree distributions have natural cut-offs. When one analyze the empirical data, it may happen that they have strong intrinsic noise due to the finite size of data-sets. For these reasons, it is advisable to measure the cumulative degree distribution $P_c(q)$ instead of $P(q)$ [35]. The cumulative degree distribution is defined as $P_c = \sum_{k=q}^{k=\infty} P(q)$. For the distributions with the power-law behavior, cumulative distribution is behaving in the similar way, i.e., $P_c(q) \sim q^{-\gamma_c}$. The exponent γ_c depends on γ as $\gamma_c = \gamma - 1$. The cumulative distribution is more appropriate because the statistical fluctuations generally present in the tails of the distribution are smoothed. Another approach, is to consider logarithmically binned data [81]. We use both of these method in analyzing the empirical data from Blogs and Digg.

As it was stressed, the complex systems ca be represented with networks of different type, depending on their nature. For this reason, the extension of some of the topological concepts is needed. For instance, in directed networks one can distinguish between incoming and outgoing links, meaning that for every node we can define two degrees in- (q_{in}) and out-degree (q_{out}). For networks of this type we consider two degree distributions, $P(q_{in})$ and $P(q_{out})$, and they may differ (see for instance Ref. [18, 12, 16, 17].

In weighted networks beside node degree, one can also define *node strength* as sum of weights of links adjacent to node i, $l_i = \sum_j W_{ij}$. The strength distribution is also used as a measure of network heterogeneity on the level of the nodes [17]. Degree

and strength, can be both indicators of the level of involvement of a node in the surrounding network. For description of weighted networks one needs to consider both of these properties. For instance, a node i can have only few neighbors to whom it may be linked with links of high weight. Its influence on these few nodes is strong, but the depth of its actions in the network network is small. On the other hand, node j can have the same strength but is weakly connected to large number of nodes. It has higher degree and thus stronger influence on the network.

Mixing patterns

The degree and strength distributions determine the statistical properties of uncorrelated networks. However, a large number of real networks are correlated in the sense that the probability that a node with degree q is connected to another node with say degree $q^{'}$, depends on $q^{'}$. The correlations between degrees are measured with *conditional probability*, $P(q^{'}|q)$, which represents the probability that link from node with degree q points to a node of degree $q^{'}$. This function satisfies the normalization condition $\sum_{q'} P(q|q^{'}) = 1$ and detailed balance condition $qP(q^{'}|q)P(q) = q^{'}P(q|q^{'})P(q^{'})$. For uncorrelated networks, detailed balance and normalization condition give $P(q^{'}|q) = \frac{q^{'}P(q^{'})}{<q>}$.
In finite size empirical networks the evaluation of conditional probability is hard, due to extremely noise data [35, 80]. For this reasons another measure for degree correlations is used, *the average nearest neighbors degree* of a node i, $q_{nn,i}$

$$q_{nn,i} = \frac{1}{q_i} \sum_j A_{ij} q_j \ . \tag{2.2}$$

By averaging $q_{nn,i}$ over all nodes with degree q, $< q_{nn} >$, one obtains an expression that implicitly incorporates the dependence on q [82]. This quantity is connected with conditional probability as

$$< q_{nn} >= \sum_{q^{'}} q^{'} P(q^{'}|q) \ . \tag{2.3}$$

For uncorrelated networks Eq. 2.3 gives $< q_{nn} >= \frac{<q^2>}{<q>}$, i.e. $< q_{nn} >$ is independent of q. If $< q_{nn} >$ is an increasing function of q, then the network is *associative*, otherwise, if $< q_{nn} >$ decreases with q the network is reffed as *disassortative* [83]. For assortative networks it is characteristic that nodes tend to connect to nodes with similar degree, while in disassortative networks nodes with low degree are more likely to be connected to ones with high number of neighbors. Assortative mixing is characteristic for social networks, where humans tend to socialize with ones with similar number of connections, while disassortativity is more characteristic

for technological networks [35].
In weighted networks one can consider *strength-strength* correlations alongside one related to degree, by investigating dependence of the average strength of the nearest neighbors of node with strength l, $< l_{nn} >$, as function of l. Following the Eq. 2.2 the *average strength of the nearest neighbors of node i* is defines as

$$l_{nn,i} = \frac{1}{q_i} \sum_j W_{ij} l_j \ , \tag{2.4}$$

and $< l_{nn} >$ as

$$< l_{nn} >= \frac{\sum_i l_{nn,i} \delta_{l_i,l}}{S} \ , \tag{2.5}$$

where S represents the number of nodes in the network with the strength l. Weighted assortative networks are the one where $< l_{nn} >$ is increasing function of l. This means, that nodes of the similar strength are more often connected among each other.
One can also consider the *weighted average nearest neighbors degree* of node i defined as

$$q^w_{nn,i} = \frac{1}{l_i} \sum_j W_{ij} q_j \ . \tag{2.6}$$

This is the local weighted average of the nearest neighbor degree, according to the normalized weight of the connecting edges $\frac{W_{ij}}{l_i}$. Such a quantity allow to characterize the assortative/disassortative behavior in weighted networks which incorporates both degree and strength. When $q^w_{nn,i} > q_{nn,i}$ the edges with larger weights are pointing to the neighbors with larger degree, while $q^w_{nn,i} < q_{nn,i}$ is the opposite case. The $q^w_{nn,i}$ thus measures the effective affinity to connect neighbors with high or low degree according to the strength of the actual interactions. Similarly, the behavior of $< q^w_{nn,i} >$ indicates the assortative/disassortative properties considering the actual interactions among the system elements.
In bipartite networks, degree or strength correlations show the connectivity patterns among nodes of different type. For instance, lets consider the undirected bipartite network with links of equal weights and with nodes of type A and B. By definition, in this networks the first neighbors are nodes of other type, so the average nearest neighbor degree is defined as

$$q^{B(A)}_{nn,i} = \frac{}{q^{A(B)}_i} \sum_j W_{ij} q^{B(A)}_j \ . \tag{2.7}$$

The growth of function $q^{B(A)}_{nn,i}$ with increasing of $q^{A(B)}$ indicates that nodes of type A are connected to nodes B which have similar degree. For example, in the case of bipartite movie networks from $IMDb$ [84], $< q^M_{nn} >$ and $< q^U_{nn} >$ (where M/U

denotes movie/user nodes) decrease with q^U and q^M respectively. This indicates a disassortative mixing in both user and movie partitions. This indicates that users with small activity tend to comment very popular movies.
For the case of directed network, one needs to consider correlations between in- and out-degree/strength. These networks can have different mixing patterns for different choice of degree/strength pairs.

2.1.2 Eigenvalue spectral analysis method for community structure

In real network, the degree (strength) distribution is broad showing high inhomogeneity among its nodes. Distribution of edges in the network is also often inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups. This feature of complex networks is called *community structure* or *clustering* [85]. *Communities*, also called *clusters*, *modules* or *subgraphs*, are groups of vertices which have common properties and/or play similar roles within the graph. There is no unique definition of network community. In many works, communities are natural, non-overlapping subgraphs of networks. Other studies consider networks in which communities overlap or are hierarchically embedded into one another. The communities identification is a remarkably attractive research direction in complex network science. The main aims of studies is to find natural groups of nodes and connections between them, to uncover and understand the structure of a network and the system it represents.
Depending on network properties and definition of subgraphs, different methods and algorithms for detection of community structure are developed [53, 44, 45, 50, 51, 53, 86]. Mostly these methods use the centrality measures, i.e., a topological [44] or a dynamical [45] flow based on the maximal-flow-minimal- cut theorem [46]. Further effective approaches for graph partitioning utilize the statistical methods of maximum likelihood [50, 51], occurrence of different time scales in the dynamic synchronization [53] and eigenvector localization [86] in mesoscopically inhomogeneous structures. In more formal approaches, the definitions of different mesoscopic structures in terms of simplexes and their combinations, simplicial complexes, are well known in the graph theory [87]. This approach has been recently applied [88] to scale-free (SF) graphs and some other real-world networks.
Properties of the eigenvalues and eigenvectors of the adjacency matrix of a complex network and of other, e.g., Laplacian matrices related to the network structure, contain important information that interpolates between the network structure and dynamic processes on it. One of the well-studied examples is the synchronization of phase-coupled oscillators on networks [35, 53, 89], where the smallest eigenvalue of the Laplacian matrix corresponds to the fully synchronized state. The synchro-

nization between nodes belonging to better connected subgraphs, *modules*, occurs at somewhat smaller time scale [53, 90] corresponding to lowest nonzero eigenvalues of the Laplacian. The positive/negative components of the corresponding eigenvectors are localized on these modules [35, 86]. The spreading of diseases [91] and random walks [42] are other types of the diffusive processes on networks, which are related to its structure and can be used for identification of communities.
In this work we define community as a group of nodes which are more densely connected to each other than to the rest of the network. The networks of on-line communication are very dense, have high density of connections. Most of the listed methods are not appropriate for densely connected graphs. For these reasons we developed two methods which are suitable for this kind of networks. One is *maximum likelihood method*, which uses machinery of mixture models and numerical technique known as the expectation-maximization algorithm for finding subgraphs in directed and undirected binary and weighted networks [51]. The other method is based on eigenvalue spectral analysis of Laplacian matrix, which is efficient for finding subgraphs in both weighted and unweighted undirected networks. Since maximum likelihood method has number of groups as input parameter we use the eigenvalue spectral analysis method when such parameter is not available.
In this Section we give detailed description of the eigenvalue spectral analysis method and apply it to networks of known community structure. First we describe the model by which we grow modular networks and then show how eigenvalue spectral analysis method can be used for identification of the communities in these networks.

Growing modular network

Real world networks often exhibit community structure with modules of different sizes and structure. In order to test the efficiency and accuracy of different methods for community detection for real networks, one needs a model of networks with similar structural properties. Our model of modular networks [55], is based on the model for growing clustered scale-free graphs originally introduced in Ref. [7]. The preferential attachment and preferential rewiring during the graph growth leads to the correlated scale-free structure, which is statistically similar to the one in real WWW [7]. Two parameters, α and M as explained below, fully control the emergent structure. Here we generalize the model in a nontrivial manner by allowing that a new module starts growing with probability P_0. The added nodes are attached preferentially within the currently growing module, whereas the complementary rewiring process is done between all existing nodes in the network. The growth rules are explained in detail below. At each time step t we add a new node i and M new links. With probability P_0 a new group, *module*, is started and the current group index is assigned to the added node (first node belong to the first group). The group index plays a crucial role in linking of the node to the rest of the network. Note that each link is, in principle, directed, i.e., emanating from the origin node and

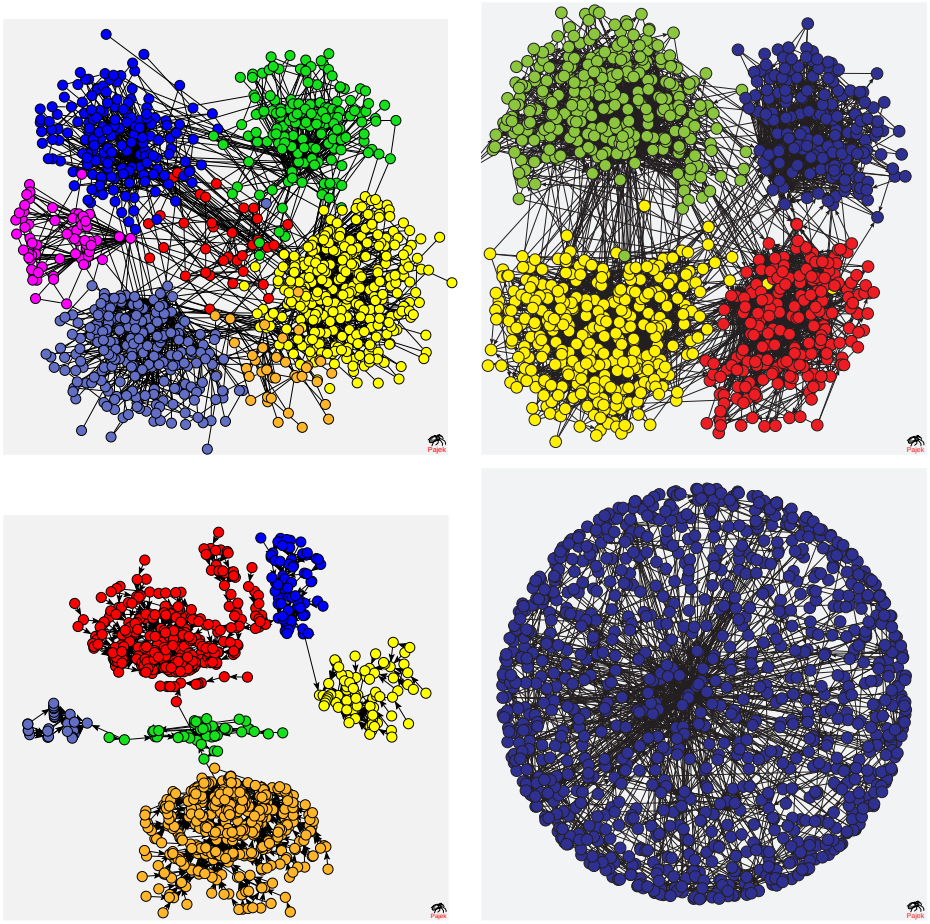

Figure 2.2: Examples of modular graphs with $N = 1000$ nodes and $M \times N$ links grown from the model rules for different values of the control parameters: (top left) $M = 2$, $P_0 = 0.006$, $\alpha = 0.9$ ($Net269$); (top right) $M = 5$, $P_0 = 0.004$, $\alpha = 0.8$ ($Net548$); (bottom left) Modular scale-free tree $M = 1$, $P_0 = 0.006$ and $\alpha = 1$ ($Net161$); (bottom right) Non-modular network $M = 1$, $P_0 = 0$ and $\alpha = 0.25$ which has degree distribution similar to one obtained for WWW [7]. Color scale of nodes for modular networks indicates their group index.

pointing to the target node. For each link the target node, k, is always searched within the currently growing module. Group membership of node is given by its group index g_k, i.e., $g_k = c$ denotes that node k belongs to community c. The target is selected preferentially according to its current number of incoming links $q_{in}(k,t)$. The probability $p_{ij}(k,t)$ is normalized according to all possible choices at time t,

$$p_{in}(k,t) = \frac{M\alpha + q_{in}(k,t)}{MN_{g_k}(t)\alpha + L_{g_k}(t)} \ , \tag{2.8}$$

where $N_{g_k}(t)$ and $L_{g_k}(t)$ stand for, respectively, the number of nodes and links within the growing module g_k. The link $i \to k$ is fixed with the probability α. If $\alpha < 1$, there is a finite probability $1 - \alpha$ that the link from the new added node $i \to k$ is cut "rewired" and a new origin node n is searched from which the link $n \to k$ established and fixed. The new origin node n is searched within all nodes in the network present at the moment t. The search is again preferential but according to the current number of outgoing links $q_{out}(n,t)$ [7],

$$p_{out}(n,t) = \frac{M\alpha + q_{out}(n,t)}{MN(t)\alpha + L(t)} \ , \tag{2.9}$$

where $N(t) = t$ and $L(t) \times N(t)$ are total number of nodes and links in the entire network at the moment t. Note that the number of added links is smaller than M for the first few nodes in the module until $M - 1$ nodes are in the module. We are interested in sparse networks, for instance, $M = 2$, the second added node in a new module can have only one link pointing to the first node in that module. The second link is attempted once within the rewiring procedure the probability $1-\alpha$, otherwise it is not added. It is also assumed that nodes have no incoming or outgoing links when they are added to the network; i.e., $q_{in}(i,i) = q_{out}(i,i) = 0$. Some examples of the emergent modular graphs of size $N = 1000$ nodes are shown in Fig. 2.2.

The structural properties of networks grown with this model depend crucially on three control parameters: the average connectivity M, the probability of new group P_0, and the attractivity of node α. By varying these parameters we control the internal structure of groups (modules) and the structure of the network connecting different modules. Here we explain the role of these parameters. Note that for $P_0 = 0$ no different modules can appear and the model reduces to the case of the clustered scale-free graph of Ref. [7] with a single giant component. In particular, for $M = 1$ and $P_0 = 0$ and $\alpha < 1$ the emergent structure is clustered and correlated scale-free network Fig. 2.2 (bottom right). For instance, the case $\alpha = \frac{1}{4}$ corresponds to the statistical properties measured in the WWW with two different scale-free distributions for in- and out-degrees and nontrivial clustering and link correlations (disassortativity) [7]. On the other hand, for $M = 1$ and $P_0 = 0$, $\alpha = 1$ a scale-free tree is grown with the power-law in-degree with the exponent $\tau = 3$ exactly.
Here we consider the case $P_0 > 0$, which induces different modules to appear statistically. The number of distinct groups (modules) is given by $G \sim P_0 N$. By varying the parameters M and α appropriately, and implementing the linking rules as explained above with the probabilities given in Eqs. 2.8 and 2.9, we grow the modular graphs with G connected modules of different topology. In particular for $\alpha < 1$ the scale-free clustered and correlated subgraphs appear cf. Fig. 2.2 (top left). Whereas for $\alpha = 1$ the emergent structure is a tree of scale-free trees if $M = 1$ (Fig. 2.2 (bottom left)). Another limiting case is obtained when $\alpha = 1$ and $M \geq 2$, resulting in a

scale-free tree connecting the unclustered uncorrelated scale-free subgraphs.
Note that the growth rule as explained above leads to a directed graph with generally different connectivity patterns for incoming and outgoing links. Each module also tends to have a central node (local hub) through which it is connected with the rest of the network. The pattern of directed connections of the nodes within modules and the role of the connecting node can be nicely seen using the maximum likelihood method for graph partitioning, as shown in our previous work [51]. For the purpose of the present work, in this paper we analyze undirected binary graphs, which have symmetric form of the adjacency matrix and the normalized Laplacian matrix. Therefore, the total degree of a node $q = q_{in} + q_{out}$ is considered as a relevant variable for which we find a power-law distribution according to

$$P(q) \sim q^{\tau} \ . \tag{2.10}$$

In Fig. 2.3 we show the ranking of nodes according to their degree for two networks, which are shown in Figs. 2.2 (top and bottom left). The ranking distribution appears to be broad (Zipfs law) with the exponent γ, which is related to the exponent in Eq. 2.10 with a general scaling relation,

$$\tau = \frac{1}{\gamma} + 1 \ . \tag{2.11}$$

The points in the flat parts of the curves at large connectivity represent the module

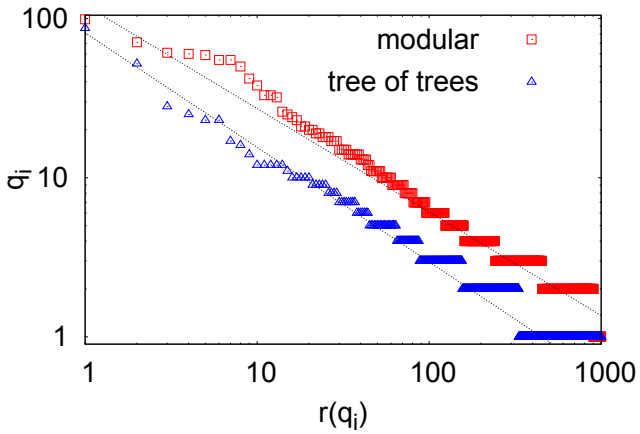

Figure 2.3: Ranking distribution (Zipf's law) of nodes according to the total node degree q_i for networks shown in Fig. 2.2 (top left) and 2.2 (bottom left).

hubs, which appear to have similar number of links. In the case of $Net269$ there are

about six such nodes, whereas in the case of tree of trees ($Net161$) two nodes, hubs of the largest subgraphs, are separated from four other hubs, and then the rest of nodes. The occurrence of local hubs changes the overall slope of the curve, compared to the networks without modules, where one finds analytically $\gamma = \frac{1}{1+\alpha}$ and thus $\tau = 2 + \alpha$ [7, 92]. Here we have approximately $\tau \approx 0.65$ for modular network $Net269$, and $\gamma \approx 0.72$ for tree of trees, $Net161$. According to Eq. 2.11, $\tau \approx 2.5$ and $\tau \approx 2.4$, for these two networks, respectively, suggesting how the modularity affects the degree distribution.
The networks which we consider in this Section to show properties of eigenvalue spectral analysis method are $Net269$ and $Net548$ (shown in Fig. 2.2 (top)). The network $Net269$ grown with direct implementation of the above rules with parameters $M = 2$, $P_0 = 0.006$, which gives $G \approx P_0 N \geq 6$ distinct modules, and $\alpha = 0.9$, leading to 10% links rewired. The second network we consider, $Net548$, is more dense $M = 5$ and has less modules $G = 4$ ($P_0 = 0.004$) with 20% of rewired links. The nodes within the modules are more densely connected compared to the rest of the networks for both $Net269$ and $Net548$, but since the overall connectivity is smaller in $Net269$ the modules are less distinguishable than in $Net548$.

Eigenvalue Spectral Analysis Method

The sparse network of size N is defined with an NxN adjacency matrix A with binary entries $A_{ij} = (1, 0)$, representing the presence or the absence of a link between nodes i and j. For the sparse binary networks the eigenvalue spectral density of the adjacency matrix is qualitatively different from the well-known random matrix semicircular law [93]. Moreover, in a large number of studies it was found that the eigenvalue spectra differ for different classes of structured networks [35]. It was showed in Ref. [55] that spectral properties of Normalized Laplacian defined as

$$L_{ij} = \delta_{ij} - \frac{A_{ij}}{\sqrt{(q_i q_j)}} \ , \tag{2.12}$$

can be used for identifying community structure in the networks. Here δ_{ij} is defined as Kronecker delta. The spectrum of Laplacian 2.12 is bounded within the interval $[0, 2]$, regardless of the size of network. The minimum eigenvalue $\lambda = 0$ always exists, and it is non degenerate if the graph consists of one connected component. In the case of graphs with communities each of the modules tends to behave as an independent network with its own zero eigenvalue.
Owing to the weak coupling between these subnetworks, we find one zero eigenvalue and a number of small eigenvalues $0 \lesssim \lambda$ corresponding to the number of topologically distinct modules. A typical spectral density of L of an ensemble of networks with six modules and the average connectivity $M \geq 2$ shows the extra peak at small

eigenvalues, as shown in Fig. 2.4 (middle) and (bottom) . In the sparse graphs and particularly in trees, the nodes with least number of links $q_m = 1$ and $q_m = 2$ play a special role in the form of the spectrum [93] near the sharp peak in the adjacency matrix at $\lambda^A = 0$.
In the network with modular stricture, $Net269$, Fig. 2.2 (top left), we have 10% of rewired links, which leave as much of the nodes with $q_m = 1$. Consequently, the central peak occurs, as shown in Fig. 2.4 (left). The presence of cycles, however, leads to the two symmetrical peaks as well as the extra peak at small eigenvalues due to the presence of modules, Fig. 2.4 (right). Comparison of the spectral densities in Figs. 2.4 (left) and (right), suggests that the increase of the minimal connectivity of nodes while keeping the same number of topological modules, the central part of the spectrum approaches the one of a random binary graph (with disappearing central peak) and a gap opens between the lower and central part of the spectrum.

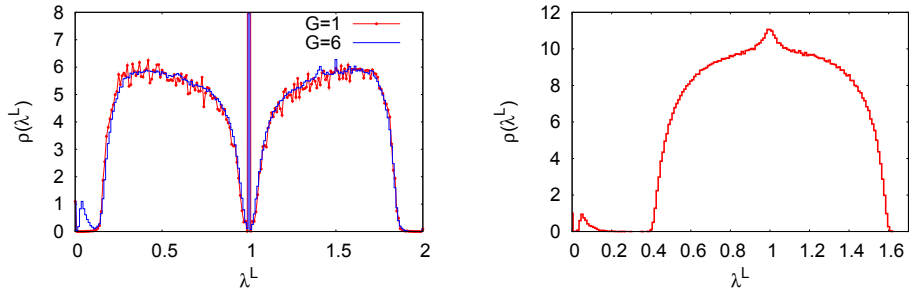

Figure 2.4: Spectral density of the normalized Laplacian for network with average connectivity $M = 2$ and $\alpha = 0.9$ (left). Spectral density of the normalized Laplacian for scale-free network with $G = 6$ modules and average connectivity $M = 5$ and $\alpha = 0.9$ (right). In each case the network size is $N = 1000$ nodes and averaging is taken over 750 network samples.

For each eigenvalue λ_i^L, $i = 1, 2, \cdots, N$, we have an associated eigenvector $V(\lambda_i^L)$ with the components V_κ, $\kappa = 1, 2, \cdots N$. A *localization* implies that the *nonzero components* $V_\kappa \neq 0$ *of the eigenvector coincide with a particular set of geometrically distinguished nodes on the network*. Specifically, for the case of the graphs $Net269$ and $Net548$, the eigenvectors associated with the lowest nonzero eigenvalues appear to be well localized on the network modules, as shown in Fig. 2.6. The origin of such localization of the eigenvectors corresponding to the lowest eigenvalues has been discussed in the literature [86], and it is related to the property of the Laplacian. The eigenvector corresponding to the trivial eigenvalue $\lambda^L = 0$ for the connected network has all positive components and κ^{th} component is proportional to $\sqrt{q_\kappa}$. When networks consists of G disconnected subgraphs, each of G eigenvectors with $\lambda^L = 0$

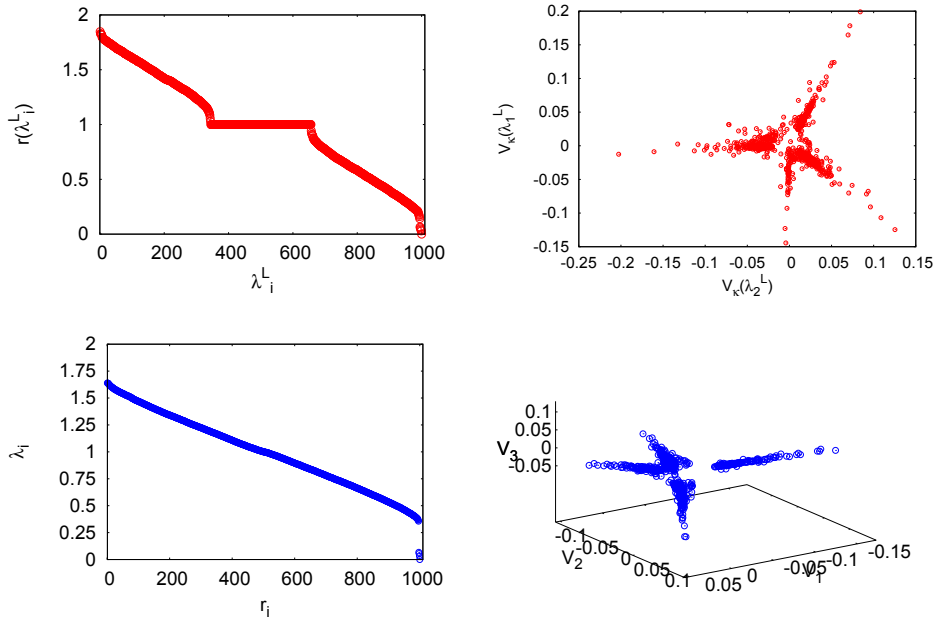

Figure 2.5: Ranking of the eigenvalues (left) and scatter plot of the eigenvector components for the eigenvalues: (top) $\lambda_1^L = 0.047033$ and $\lambda_2^L = 0.038286$ for the network $Net269$ and (bottom) $\lambda_1^L = 0.066962$ and $\lambda_2^L = 0.055717$ and $\lambda_3^L = 0.033017$ for the network $Net548$.

has non-zero components only for nodes within one module. If the subgraphs are not fully disconnected, but instead, few links exist between them, the degeneration of the zero eigenvalue disappears, leaving only one trivial eigenvector and $G - 1$ approximately linear combinations of the eigenvectors of the modules. For the or-

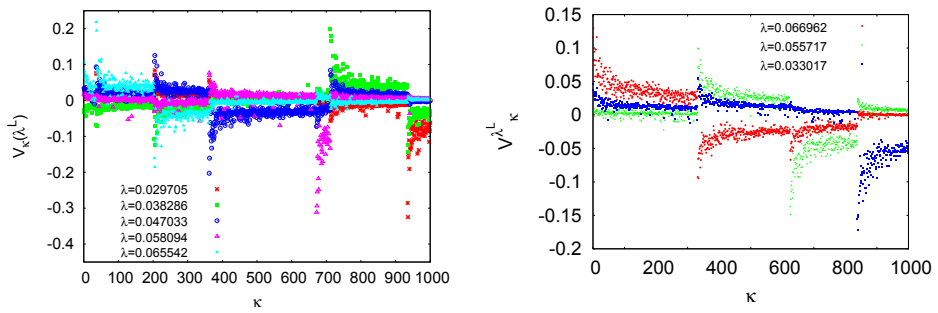

Figure 2.6: Eigenvector components, indicated by colors scale, for five $Net269$ (left) and three $Net548$ (right) lowest non-zero eigenvalues of the normalized Laplacian.

thogonality reasons, these linear combinations have components of both signs, as opposed to all positive components of the $V(\lambda^L = 0)$ vector. In the case of well separated modules, the components corresponding to one subgraph appear to have the same sign. The more links between subgraphs exist, the distinction between modules appears fuzzier. The structure can be also seen in the *scatter plots* in Fig. 2.5(right) where the eigenvector components belonging to two small eigenvalues are plotted against each other. In this projection each point corresponds to the index of one node on the network. The separated branches contain the indexes of the nodes belonging to different modules. The greater the differentiation between the modules the less common points of the branches. The communities in networks $Net548$ and $Net269$ are clearly separated which is manifested as empty space around origin in scatter plots, Fig. 2.5 (right). The existence of the points close origin to suggest better interconnection between groups. Note that one can consider scatter plots in two (see Fig. 2.5 (top right)) or three (see Fig. 2.5 (bottom right)) dimensions by taking into account the corresponding number of eigenvectors.
In this thesis we use C++ code for finding eigenvalues and eigenvectors of symmetric normalized Laplacian matrix. For details see Appendix A.

2.2 Statistical analysis approach

The considered empirical data have high resolution in time, i.e. time of every event is recorded. This enables quantitative study of the dynamics of the system trough different kind of quantities. We study temporal patterns of user behavior and activities on the posts for better understanding of behavior of humans in on-line social communities. Analysis of time series of users activity reveals temporal correlations and evidence of self-organized criticality in dynamics of on-line social communities.

2.3 Temporal patterns

Patterns, temporal or spatial, are characteristics of a system and therefore indicators of essential underlying processes and structures. They contain information about the internal organization of the system, but in a wrapped form. Analysis of patterns and their formation is useful for understanding and predicting the behavior of complex systems.
Activity of elements in complex systems is recorded in time and can be represented as temporal activity pattern (see examples in Fig. 2.7). On the y axe are shown indexes of system elements, while x is time axe. Further, activity of an element is represented with a point in XY plane. Depending on system properties and behavior, these temporal pattern can be *regular* or *fractal*. Fractal structure means that

certain temporal patterns repeat over different time scales. The fractality nature of temporal and spatial patterns was found for many complex systems such as stock markets [94], human behavior[12, 16, 18, 95, 96], or World Wide Web [97]. The Fig. 2.7 (left) shows temporal patterns of more than 3000 users on B92 Blog [12]. Pattern shows high heterogeneity in user's activity on the Blog. The fractal temporal pattern of users activity is manifested on events recorded at Blogs, in particular, each user action results in either a new post or a new comment to some of the existing posts. From the data one can trace where each particular user action was directed to, by analysis of the temporal patterns of the related posts. The temporal pattern of linking to a particular post is shown in Fig. 2.7 (right) for first 250 posts on B92 Blogs. It shows that posts becomes less interesting after certain period of time. We will show that similar pattern is observed in the case of other Blog sites.

The fractality in temporal pattern can be further quantified by analysis of time

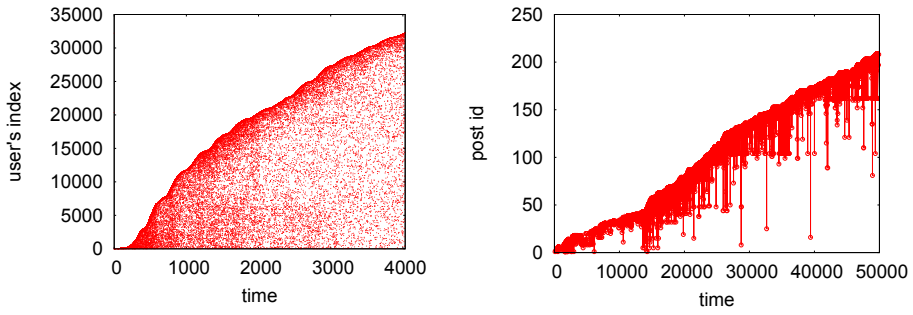

Figure 2.7: (left) Temporal linking pattern of users (ordered by time of first appearance) within roughly first year since the opening of the B92 Blogs. (right) Linking pattern over times on B92 posts.

interval Δt between two successive user activities or between two events on the post. The distribution of time intervals between two successive activities can be a good indicator of the process type. For instance, in the random process the time distribution between two successive activities is normal distribution with some characteristic time $< t >$. On the other hand, different studies on different social systems show that these distributions have power-law behavior of type $\frac{1}{\Delta t^{\alpha}}$ [12, 16, 17, 18]. Similar studies behavior was found for distribution of waiting times between cause and action performed by humans. The appearance of the power-law distribution of the response times $P(t - t_i)$ to an event i posted at time t_i was associated with human nature of acting with priority queues [95, 96, 98]. Assuming a single-server-queue limit and random arrival of events to the queue, in reference [96] a universal power-law distribution of the waiting (or response) times was derived with the exponent 3/2, corresponding to the situation when the average arrival rate exceeds the average execution rate. An exponent larger than 2 is expected in the opposite situation,

as discussed in reference [95], where attention rather than priority was a key mechanism. In both cases, a theory of independent queues were considered. Note that in the network environment, where the queues are mutually interacting, for instance in the packet queuing processes [99] with LIFO queue, the waiting time distributions exhibit power-law with exponents depending of the traffic density, dropping below 2 when the jamming on the network occurs.

2.4 Time-series analysis

Beside temporal patterns, in this thesis we will also consider analysis of time series of activities. A time series is defined as a sequence of observations ordered in time. Mostly these observations are collected at equally spaced, discrete time intervals. The analysis of time series help us to identify the nature of the phenomenon represented by the sequence of observations and to predicting future values of the time series variable. In on-line communication on Blogs and similar websites, the activity is expressed as posting of comments and messages. The temporal evolution of number of comments or messages is considered as a signal from our system. Depending on the level we want to study the system dynamics, we can consider different kinds of time series:

- *Cumulative time series* of the number of comments posted within the whole system or in one community Fig 2.8 (left). The y axes shows the number of comments posted within one time bin.

- *Time series* of activity of single user or activity on the single post Fig. 2.8 (right), where y axes shows the number of comments posted by user on post during one time bin.

Several quantities, which can serve as indicators of type of dynamics in the system, can be calculated from these time series.

Taylor's fluctuation scaling

Complex systems consist of many interacting elements which participate in some dynamical process. The activity of various elements is often different and the fluctuation in the activity of an element grows monotonically with the average activity. This relationship is often of the form

$$\sigma_X \approx C \times < X >^{\mu} \tag{2.13}$$

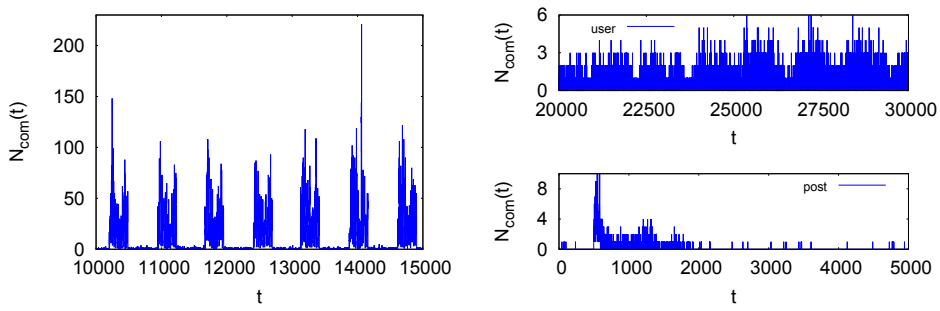

Figure 2.8: (left) Part of time series of number of comments in B92 Blog. (right) Examples of time series of number of comments on post (bottom) and written by one user (top) on Digg. All time series are calculated for time bin equal to one hour.

where σ_X represents standard deviation of X and $< X >$ its average. The exponent μ is in the range $[1/2, 1]$. This power law has been observed in a wide range of systems, ranging from population dynamics through the Internet to the stock market and it is often treated under the names *Taylor's law or fluctuation scaling*. The largest exponent $\mu = 1$ is generally expected in strongly driven dynamical system, whereas the lower limit $\mu = \frac{1}{2}$ is found in random (uncorrelated) events [100]. In the case of blogging dynamics one can consider Taylor's scaling by calculating the dependence of the fluctuations σ on average $< N_{com} >$ value of time series for single user or all users on single post. This dependence is visualized as scatter plot with $< N_{com} >$ on x axes and σ on y axes, where every time series is represented as a point in XY plane [12].

Power spectrum

One way to consider the signal from some system, is in the discrete time domain, which puts a series of values consecutively in time. By analyzing this signal we can investigate system activity at every moment in time, and can also make some statements about its long-term behavior. However, it is rather difficult to say anything about how the long-term behavior is related to the short-term development of the signal. Another way to look at a signal is to view its spectral density (i.e., the Fourier transform of the signal). Fourier analysis is the study of how general functions can be decomposed into trigonometric or exponential functions with definite frequencies [101]. The Fourier transform swaps the dimension of time with the dimension of frequency. One can think of the Fourier transform as a combination of slow and fast oscillations with different amplitude. A very strong and slow component in the frequency domain implies that there is a high correlation between the

large-scale pieces of the signal in time (macro-structures), while a very strong and fast oscillation implies correlation in the micro-structures. Therefore, if our signal $f(t)$ represents values in every single moment of time, its Fourier transform $F(\nu)$ represents the strength of every oscillation in a holistic way in that chunk of time. These two signals are related to each other according to formula:

$$F(\nu) = \int_{-\infty}^{+\infty} f(t) \exp(-i\nu t) dt \ . \tag{2.14}$$

In the Fourier transform, oscillations are characterized with sinusoid functions. The average value of any smooth oscillation, fast or slow, strong or weak, is zero. The power of these oscillations can be studied in the same way as the original signal. Parseval's theorem for energy signals states that:

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |F(\nu)|^2 d\nu \ . \tag{2.15}$$

The Fourier transform analysis assumes the life of a signal from $-\infty$ to ∞. For that reason when an analysis is carried out for a finite amount of time, it is either assumed that the signal is periodic or that it has a finite amount of energy. A true power spectrum of a signal has to consider the signal from $-\infty$ to ∞. However, we are not always able to observe a signal that way or derive precise functions for it. We can define $F_T(\nu)$ which is the Fourier transform of the signal in period T, and define the power spectrum as the following:

$$S(\nu) = \lim_{T \to \infty} \frac{1}{T} | \ F_T(\nu) \ |^2 \ . \tag{2.16}$$

According to the Wiener-Khinchin theorem, $S(\nu)$ is the Fourier transform of the autocorrelation function, $C(\tau)$, i.e.,

$$C(\tau) = \frac{1}{2\pi} \int_{0}^{\infty} S(\nu) \exp(i\nu t) d\nu \ . \tag{2.17}$$

Auto-correlation function represents the relationship of long and short-term correlation within the signal itself, and it is defined as

$$C(\tau) = \int_{+\infty}^{-\infty} f(t+\tau) f(t) dt \ . \tag{2.18}$$

A relationship between different sections of the power spectrum implies a relationship between the auto-correlation of the signal in the time domain to which those frequency sections are referring. Signal from the systems without any correlations, so called "white noise" signals, are random. The power spectrum of these signals

is homogeneously distributed across all frequencies. A large number of real systems has power-law behavior of power spectrum, $S(\nu) \sim \frac{1}{\nu^\beta}$ within some interval of frequencies. When $\beta \simeq 1$ the system exhibits long range temporal correlations, and it signal is often reffed as "pink noise". The $\frac{1}{\nu}$ present behavior which is strongly influenced by its entire history. Its memory is dynamic; the influence of recent events is added to and gradually supersedes the influence of distant events. The influence of the distant past fades very slowly. Often these systems are found in *self-organized critical state* [65]. As we will show in the next chapter, the user's activity on the Web has similar properties, indicating that complex systems of techno-social interactions may be found in self-organized critical state.

Avalanches

The time signal from different dynamical system are often observed to be composed of apparently distinct bursts or pulses, which in the limit of slow driving are associated with distinct and typically spatially localized avalanches of activity occurring in the system. A typical feature of such bursts or avalanches is that the statistics of various measures associated with them appear to lack a characteristic scale, e.g. the avalanche sizes and also durations are usually characterized by power law distributions.

In the complex systems such as earthquakes [102, 103] or Barkhausen noise [104], the avalanches can be readily determined from the measured time series. Specifically, putting a baseline on the level of random noise, an avalanche encloses the connected portion of the signal above the baseline. The size of the avalanche, S, is therefore calculated as surface area limited by the signal between two points in of its intersection with baseline. Beside avalanche sizes one can also consider the two temporal properties of bursting behaviors: time interval between two consecutive avalanches δT and the time-life of avalanche ΔT. Distributions of both these values also exhibit power-law behavior for the systems in SOC state.

Depending on system behavior one can identify avalanches of different size and duration. In *subcritical* regime the avalanches that develop are of relatively small size and do not spread trough whole system. This kind of behavior is related to non-clustered activity and obtained distribution of avalanche sizes and duration are of exponential type. On the other hand, supercritical behavior is associated with high degree of synchronized activity among elements of the system. The most of the obtained avalanches are of order of the system size which results in power-law type of distribution of avalanche sizes with superimposed peak around the avalanches of network size. The pure power-law means that there is no avalanche of characteristic size and the system is in self-organized critical state.

Chapter 3

Techno-Social Networks from the Empirical Data of Blogs and Digg

The Internet became a central part of our communication network, in both professional and in personal life. One sends emails, calls its colleges and friends using Skype, chat with them and share its information over Facebook. The development of social media enabled humans to communicate and share information with people they do not know personally, allowing them to dramatically expand their network of social contacts. During daily activity on Web, humans constantly leave fingerprints of their activity, creating massive amounts of data. The methods described in Chapter 2, applied to this data, can be used for analysis and description of social dynamics on the Web [12, 16, 17, 18, 84].
In this Chapter we will show how the data from Blogs and similar Web portals can be mapped onto networks. We will show which topological properties of these networks are closely linked with the collective emotional behavior of humans on Blog. Using methods of statistical physics and complex networks, thus we will explore dynamical features of this behavior. The knowledge obtained using this quantitative analysis will represent the basis for developing of models of human emotional activity on the Web.

3.1 Data structure

Data which we consider in this thesis were collected from different types of Web portals where people can share information (stories, links) and exchange their opinions. Specifically we analyze the data from three different Web sites:

- *B92 Blog* - a Serbian Blog owned by Serbian news site *B92*.

- *BBC Blog* - Blog owned by *BBC*. Chosen due to the fact that it is in English.

- *Digg* - is a Web site where users submit, rate and comment some content. The user's activity on this site is very high and the comments are in English.

Common feature of these three site, very important for our analysis, is that they require users to be registered to be able to write or comment stories. As a result we have a unique ID for each user, which allows us to monitor its behavior over time.
A Blog is a places where people can share their opinion about some specific subject trough posts and comments on these posts. Unlike Blogs, Digg users are finding and posting links to a different type media (news stories, videos, music) on Digg site. Together with other users they can comment links and share opinions about links and their subject. Both, Blogs and Digg sites, have similar policies, which results in similar dynamics of the system. For this reason, these system can be analyzed and model in a similar way, [12, 16, 17, 18].
B92 Blog is on of the most popular blogs in Serbian language. We have collected all data from the beginning *27. May 2007* till *1.March 2009*. On this Blog site users are registered not just to read and comment other posts, but to write their own post. No predefined categories of post subjects are imposed, thus, the internal structure of posts emerges in a self-organized manner through user interactions on posts and comment-on-comment actions. The availability of posts is time limited to seven days (this rule was imposed after first few months of the functioning). Some of the users are upgraded to so called *VIP authors*, whose number fluctuates in time, and their recent posts are highlighted. We analyzed the posts written by all VIP users in the above mentioned period and consider all users related to these posts, which comprises of $N_U = 4598$ users, and $N_P = 4784$ posts and $N_C = 406527$ comments to these posts.
BBC Blog exists for much longer time. In order to compare the data with one from B92 Blog, we have collected the data form the same period as above, which gives $N_P = 3792$ posts, and $N_C = 80873$ comments written by $N_U = 21462$ registered users. In contrast to B92 Blogs, at the BBC Blogs both authors of posts and category of posts are predefined and fixed. The collected posts were classified into five different categories: *Music&Art*, *Business&Economy*, *Nature&Science*, *Technology* and *Sport*. The users are registered and allowed only to read and comment the posts. Accessibility of posts is not limited in time. In contrast to B92 Blogs, information about ID of a comment which is commented is not available in BBC Blogs and all comments are attributed to the original post.
Although the **Digg** represents different type of Web site, interaction of users is similar to one on Blogs, and for this reasons Digg data can be analyzed using the same methods. The activity on the Digg is much higher compared to one on B92 and

BBC Blog, for which reasons we consider data set collected for the period of of 3 months, *February till April 2009*, which consists of: $N_U = 484986$ users who posted $N_P = 1195808$ stories and $N_C = 1646153$ comments. Since we are interested in dynamics as a consequence of interaction between users, we consider just part of the data set: all comments which were posted on $N_P = 129999$ stories by $N_U = 101000$ users. Unlike BBC Blog, on Digg one is allowed to leave comment-on-comment, which gives more detailed information about interactions between users. Also categories and subcategories of the posted links are predefined.
The data structures from Blogs and Digg that we consider have high-resolution (1 min) of the temporal occurrence of the events (posting a comment) and full information about both users and posts-and-comments, as well as the full Text of the posts and comments. The user is registered on all sites with a unique ID is requested. As it was stressed different policy is practiced in storing the information, especially regarding the information about comment-on-comment. Such information affects the structure of the network. The emotional classifiers are only available for the texts written in English, for which reasons dataset for B92 Blog does not contain information about emotional content of posts and comments. The analysis of this data is conducted only in a part which do not include emotional content.

3.1.1 Other on-line social networks

Last decade led to the development of various sites, together called the social network where people can establish and maintain their social contacts.
MySpace and Facebook are networks of virtual friends, from which some are friends in the off-line world. On these sites people communicate by writing messages on own or the walls of their friends. These messages are shorter than one posted on Blogs or Digg, and often have different emotional content. Although on these Web sites communications is also mediated, trough messages, there is no post as on Blogs, but the users are communicating directly. Data from obtained from these websites are mapped onto different type of networks compared to one for Blogs and Digg. These networks are monopartite, with users as nodes and links representing exchanged messages.
Twitter is an on-line *social networking* and *microblogging* site that enables its users to send and read text-based posts of up to 140 characters, informally known as "tweets". Since it was launched in July 2006 it gained worldwide popularity very rapidly, with over 300 million users as of 2011, generating over 300 million tweets and handling over 1.6 billion search queries per day. It is sometimes described as *the SMS of the Internet*. The amount of data produced on the twitter is very high compared to Blogs and Digg. The communication is fast, but not synchronous as in the case of chat rooms. A user can either post a tweet which refers to all of his

followers, or it can specify the twitter name of the targeted persons. The tweets can be personal, for which reasons the relations resembles one on social networks, or can contain information which is more characteristic for blogging behavior. Again, the lack of the posts, points to monopartite network as appropriate choice for mapping data from this site.
Internet Relay Chat (IRC) is a protocol for real-time Internet text messaging (chat) or synchronous conferencing. It is mainly designed for group communication in discussion forums, called channels, but also allows one-to-one communication via private message as well as chat and data transfer, including file sharing. IRC chat rooms are interesting for studying of synchronous collective behavior. The communication is instant in time, which is characteristic for face-to-face communication, but is devoid of physical contact and takes place on the Web. Chat channels can be used in different purpose, friendly chat, help or exchange of experience. Some of the chat rooms are very active, and can generate a large amount of data. The communication is carried through the messages, meaning that interactions and dynamics are similar to one on twitter.
Except from network mapping, methods described in Chapter 2 can be applied for quantitative analysis of this data.

3.1.2 Data collection

As it was stated, the data were collected from three different sites:

- **B92 Blog**-*blog.b92.net*,
- **BBC Blog** -*www.bbc.co.uk/blogs/*,
- **Digg** - *digg.com*.

To collect the data sets from both Blog sites, we developed three different web-bots written in *Python* language. First script, *collect_posts.py*, was used for collecting the *HTTP* addresses of posts published in a certain time period, while second script, *collect_users.py*, was used for collecting of comments from these posts together with the information about the author of the comment, the time of posting and weather it is a comment on post or on other comment (when information is available). For the case of BBC Blog we developed the third script which was used for collecting and extracting the clean text of the comments from *html* code of evry post. The scripts, together with explanation, are given in Appendix B.
The Digg dataset was collected through the websites publicly available *API*3, which

[1] *http://apidoc.digg.com/*

allows programmers to directly access the data stored at its servers. The data contain information about stories and associated comments together with users IDs, the time when the story/comment was posted, and information weather comment was on comment or on story. The Digg also enables users to give positive/negative votes to a certain story/comment, i.e. to *digg-up*/*digg-down*. The information of number of positive and negative votes and who made them is also contained in the data, but we are not using it in our analysis.

3.1.3 The Emotion Classification

The hidden emotions in text of collected posts and comments are extracted using machine learning techniques. The extraction of emotions from the text is a classification problem. In the classification procedure the document D is assigned to one or more available classes from fixed set $C = \{c_1, c_2, \ldots, c_N\}$. The classification can be carried using different classifiers which can be categorized into two types: *supervised* and *unsupervised*. We use supervised language model classifier [16] which is explained below.

Language model classifier

To classify the text in data sets collected from BBC Blog and Digg, we used two supervised classifiers. First classifier estimates a probability for document D being objective or subjective, i.e. it assigns one of the classes $C_1 = \{obj, sub\}$. The second one determines the polarity of subjective document is negative or positive, $C_2 = \{pos, neg\}$. The final output of the classifiers is therefore one of $\{obj, pos, neg\}$ on which we assigne one of the three numerical values to every document: $e = 1$ for positive, $e = -1$ for negative and $e = 0$ for objective (neutral) documents. For both classification tasks we have utilized language model classifier [105, 106]. This classifiers find the appropriate class for the document by maximizing the posterior probability $P(c|D)$ for that given class. The best class is the maximum a posteriori (MAP) class c_{MAP}:

$$c_{MAP} = \arg\max_{c \in C} \{P(c|D)\} \tag{3.1}$$

The posterior probability is given as

$$P(c|D) = \frac{P(D|c) * P(c)}{P(D)} \tag{3.2}$$

where $P(c)$ is the priori that indicates the relative frequency of class c. The $P(D)$ doesn't influence the outcome of the classification so it can be removed from Eq.

3.2 from which it follows

$$c_{MAP} \approx \arg\max_{c \in C} \{P(D|c) * P(c)\} \ . \tag{3.3}$$

The document D can be presented as a token sequence $\{w_1, w_2, \ldots, w_n\}$. The language models operate by estimating the probability of observing document D, given class c. Based on this given the class c, the model is to estimate the probability of observing the sequence $\{w_1, w_2, \ldots, w_n\}$:

$$\begin{aligned} P(D|c) &= P(w_1|c) \times P(w_2|c, w_1) \times \ldots \times P(w_n|c, w_1, w_2, \ldots, w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|c, w_1, \ldots, w_{i-1}) \ . \end{aligned} \tag{3.4}$$

To estimate the equation 3.4 one can use n-gram approximation which assumes that the probability for appearing of token w_i in the document D depends only on the $n-1$ preceding tokens:

$$P(w_i|c, w_1, \ldots, w_{i-1}) = P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}) \ . \tag{3.5}$$

The classifier is trained on the set of documents with already defined classes of the documents. The maximum likelihood estimate of $P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1})$ during the training phase of the classifier is by counting the frequency of occurrence of the tokens sequences:

$$P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{\#(c, w_{i-(n-1)}, \ldots, w_i)}{\#c, w_{i-(n-1)}, \ldots, w_{i-1}} \ , \tag{3.6}$$

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 occurrence of sequence $w_{i-(n-1)}, \ldots, w_i$ in documents of class c during the training phase.
The problem with n-gram approximation is the their small occurrence in longer n-grams (usually longer than 3), i.e. their occurrence is not enough to provide a strong indication of class preference. To solve this problem we need to further break the estimation of the probability $P(D|c)$ to smaller n-grams using the procedure called smoothing [106], such as Laplace, Good-Turing, back-off estimators or Witten-Bell approach which will be used hear [107]. Based on this method we obtain:

$$\begin{aligned} P(w_i|c, w_{i-(n-1)}, \ldots, w_{i-1}) &= \lambda(c, w_{i-(n-1)}, \ldots, w_{i-1}) \times P(w_i|c, w_{i(n1)}, \ldots, w_{i1}) \\ &+ (1\lambda(c, w_{i(n1)}, \ldots, w_{i1}) * P(w_i|c, w_{i(n2)}, \ldots, w_{i1}) \end{aligned} \tag{3.7}$$

where

$$\lambda(c, w_{i(n1)}, ..., w_{i1}) = \frac{\#(c, w_{i(n1)}, ..., w_{i1})}{\#(c, w_{i(n1)}, ..., w_{i1}) + LW(c, w_{i(n1)}, ..., w_{i1})} \ . \tag{3.8}$$

The $W(c, w_{i(n1)}, ..., w_{i1})$ is defined as the number of extensions of the specific token sequence:

$$W(c, w_{i(n1)}, ..., w_{i1}) = |\{w_k | \#(c, w_{i(n1)}, ..., w_{i1}, w_k) > 0\}| \ , \tag{3.9}$$

and L is the *hyperparameter* of the distribution and its purpose is to provide a balance between higher and lower order of n-grams. In the case of smaller training sets where shorter n-grams are more often, the value of L is larger. Hear, the value of L is set to the value of the longest n-gram.
The pseudo code for training of the algorithm is given in the Table 1. The language model classifiers are trained on the $BLOGS06$ dataset [108, 109]. This dataset is crawled of approximately 100000 English written Blogs and has been used for 3 consecutive years by the Text REtrieval Conferences (TREC). Participants of the conference were asked to find the blog post for which they think it expresses the opinion about some movie, company, person. Then this document is given to another participant to be judged: with 1 if the opinion is objective, 2 if the document contains and explicit negative opinion about the subject, 4 if it contains an explicit positive opinion. During period of three years the participants produced the dataset which contains 34379 documents, out of which around half were said to be objective, 7930 negative and 9968 positive. This documents were used as a *gold standard* for training classifiers. Specifically, for the first phase of classification, $C_1 = \{obj, sub\}$ the documents with the label 1 were taken as objective and ones with 2 and 4 as subjective. For the second phase, only documents taken as subjective were used. The resulting classifiers have accuracy of approximately 70% for either classification task. The algorithm which classifys the document is given in Table 2

Each comment/post is considered as a document and classified separately. As a result of classification we obtain two probabilities: the probability that x is objective, $P_{obj}(D)$, and the probability that subjective document is positive $P_{pos}(D)$. The document is classified as objective, i.e. its emotional content is valued as 0, if $P_{obj}(D)$ is larger than a threshold $C_{obj} = 0.43$. Notice that this means a large threshold value for the criteria of subjectivity of texts. Further, objective text is classified as positive if $P_{pos}(D)$ is found to be larger than $C_{pos} = 0.7$, otherwise it emotional content is set to $e = -1$. 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.

Lexicon-based classifier

There are other algorithms which can be used for emotional classification of the text. One is lexicon-based classifier which aims to extract the emotional con-

Algorithm 1 Train Language Model classifier

INPUT: Documents $D = \{d_1, \ldots, d_n\}$, Classes $C = \{c_1, \ldots, c_t\}$, Function $\phi : DxC \rightarrow T, F$, 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**
  $|D_c| \leftarrow |\{d_i | d_i xc \rightarrow T\}|$
  $P(c) = |D_c|/|D|$
  $W_c \leftarrow \{w_{k-(n-1)}, \ldots, w_k | \phi(d_i, c) = T\}$
  for all $w \in W_c$ **do**
    Count $\#(c, w)$
  end for
  for all $w \in W_c$ **do**
    $P(w|c) \leftarrow$ Equation 3.7
  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 $d = w_1, w_2, ...w_n$ to be classified
$W_d\{w_{k(n1)}, ..., w_k | d = w_1, ..., w_{k(n1)}, ..., w_k, ..., w_n\}$
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 $\arg\max_{c \in C} Score(d)$

tent of a document based on some dictionary with emotional values of the certain words. This classifier doesn't require training and provides two independent ratings, one for the positive scale $C_{pos} = \{1, 2, 3, 4, 5\}$, and one for the negative $C_{neg} = \{-1, -2, -3, -4, -5\}$, where bigger absolute values indicate stronger emotion and values $1, 1$ indicate its absence. The classifier runs trough text and finds the words with the biggest positive and negative emotion, and this is given as an emotional content of the document.
The two-dimensional 5-point scale human annotation can be transformed to a binary scheme using following rules:

- All documents that have $(-1, +1)$ are considered as objective.
- All documents that have $e_+ = +3$ of higher and negative e_- equal -1 or -2 are classified as positive.
- All documents with $e_- = -3$ or smaller and positive e_+ equal $+1$ or $+2$ are classified as negative.

All other combinations are considered as error and those documents are not classified.

Circumplex map

Humans have difficulties to asses, discern, and describe their own emotions [110]. This suggests that humans do not experience or recognize emotions as isolated, discrete entities, but rather as ambiguous and overlapping experiences. Similar to the spectrum of color, emotions seem to lack the discrete borders between them [111]. In most of the studies, a subject rarely describe feeling a specific positive emotion without also claiming to feel other positive emotions. These correlations among emotions, are addressed in psychology by two dimensional models of affect with different definitions of x and y scale. The Russell's $2D$ circumplex model of emotions describes each of 28 basic emotional states with pair of two variables, valence and arousal, (v, a) [112]. The valence v, indicates weather pleasure related to an emotion is positive or negative while arousal,a, indicating the personal activity induced by that emotion (see Fig. 3.1). The emotions with high positive valence and high arousal such as astonished, excited, delighted are in the top right quadrant while, very strong but negative emotions, as angry and insulted are in the top left part of the circle. The states which are low in arousal "sad" or "relaxed" are in the lower half and their position on the x-axes depends on its polarity.
One can measure the arousal and valence in text using the affective norms for English words (ANEW) [113]. The ANEW is being developed to provide a set of normative emotional ratings for a large number of words in the English language. The participants have rated the words from the ANEW on the dimensions of pleasure and

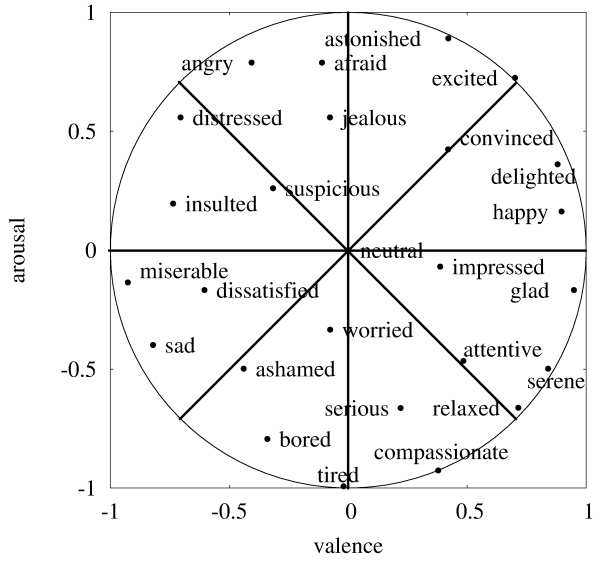

Figure 3.1: The two dimensional Russel's circumplex model of emotional states.

arousal on the scale 1 to 9. The very negative word is rated with 1 and positive with 9. In the same way the word with high arousal has value of 9.The values of arousal and valence for every word are obtained as average over single ratings given by participants. The set of (v, a) of text can be determined as average of arousal and valence of words in the text based on ANEW. The emotion obtained from the classification of textual document can be mapped onto Russell's circumplex map using following rules. First the values v and a, which are on the scale form 1 to 9, are scaled using linear transformation $0.25x - 1.25$ to obtain the values v_1 and a_1 which are in interval $[-1, 1]$. Then, the values v_1 and a_1 are mapped onto a circle according to equations

$$a_2 = \frac{a_1}{\sqrt{1+z^2}} \ , \tag{3.10}$$

and

$$v_2 = \frac{v_1}{\sqrt{1+z^3}} \ , \tag{3.11}$$

where $z = min(|v_1|, |a_1|)$ (see Refs. [114]). The obtained set of values of (v_2, a_2) for a data set, can be then binned in order to obtain distribution of emotional states. The obtained histogram is visualized as $3D$ map with valence on the x axes, arousal on y and the number of emotional state in the bin on z axes. The emotional pattern, derived in this way, gives information about emotional states that are characteristic of the users in considered techno-social network.

One of the issues with lexicon based classifier, for both negative-positive and arousal-valence scale, is small number of words in the used dictionaries. The dictionaries

especially lack of the words commonly used in on-line communication.

3.2 Structure of networks

3.2.1 Data mapping

The data collected from the Web can be mapped onto different type of networks, depending on type of interactions and on the level of information we want to keep. In on-line world, human communication is not direct but trough textual messages (comments, posts, etc.). On Blogs and similar sites, humans interact with each other by leaving comments on posts (stories). This type of data, in network terms, have naturally induced partitions, *users* and *posts&comments*, meaning that *bipartite networks* are more suitable for representation of this kind of data [12, 16, 17, 18]. Depending on th level of information one wants to keep, it can also considered different types of projected networks. The following types of networks are considered in this thesis:

- *Full bipartite network* is suitable for representation of Blog and Digg data. It consists of two partitions: users (U) and posts&comments (P + C). On Blogs user i can either read and comment some post(comment) j, denoted as link $j \to i$, or can be its author, link $i \to j$. The size of this network is $N_U + N_{P+C}$. Together with the direction of the links and times of their appearance, this network contains full information stored in the data. The mapping is illustrated in Fig. 3.2 (top left).

- Weighted bipartite network is a compressed bipartite network with users and posts as two types of nodes (U + P), while the weights of links W_{ij} represent the number of comments of the user i on the post j, as illustrated in Fig. 3.2 (top right). These networks are undirected and more suitable for the spectral analysis, compared to full bipartite networks.

- 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 onto user networks, two users are connected directly with a weighted link, representing the number of common posts per pair of users, C^{Pij} , or similarly, the number of common users per pair of posts, $C^{U}ij$, in the post-projection networks, as illustrated in Figs. 3.2 (bottom).

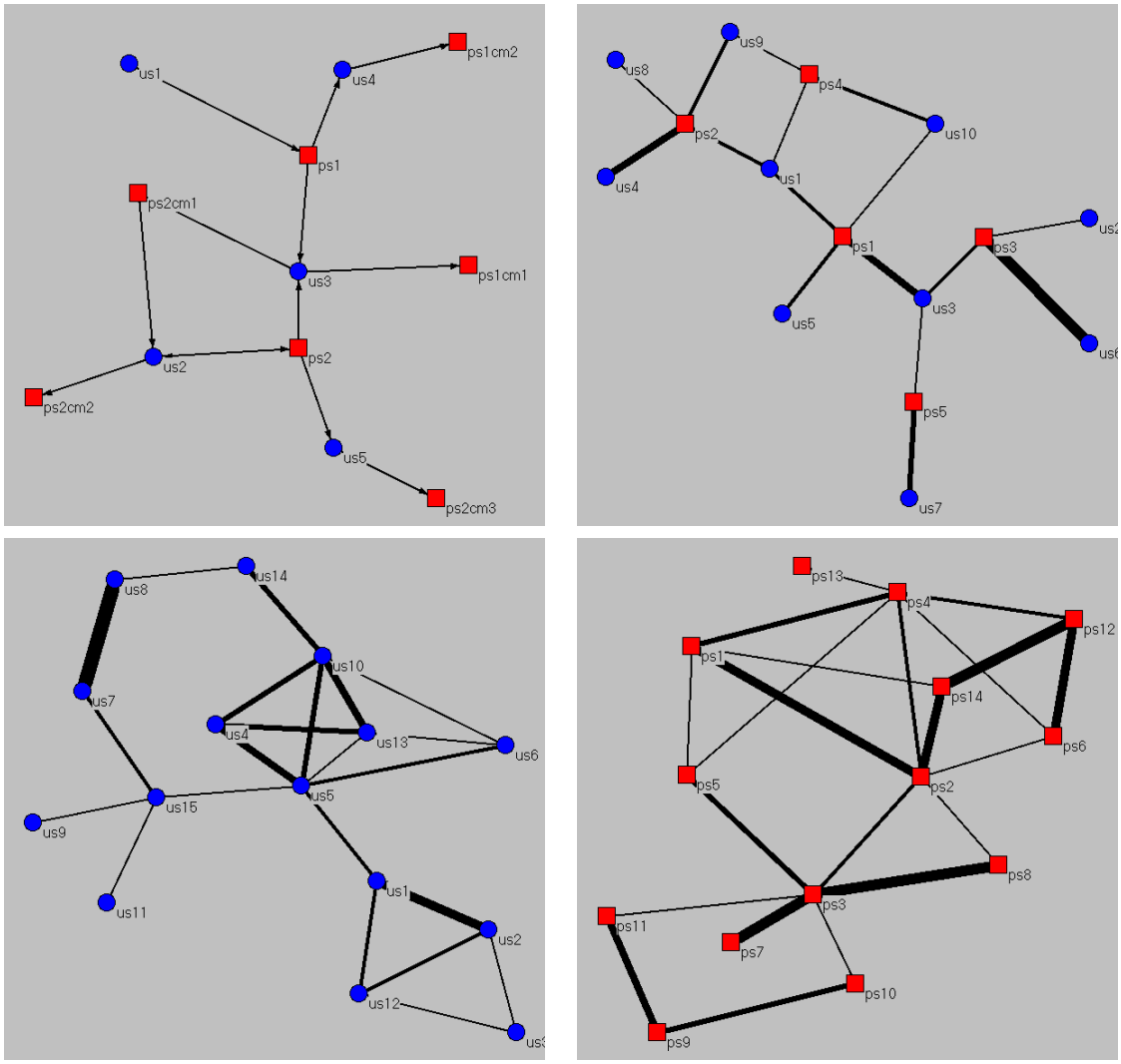

Figure 3.2: Illustration of data mapping onto a directed bipartite network with users represented by bullets, and posts and comments, represented by squares (top left), and a bipartite network of users and posts, while the weighted links represent the number of related comments (top right). Weighted monopartite networks of users connected via number of common posts (bottom left) and network of posts connected via number of common users (bottom right)

This data representation is suitable for the users in the Cyberspace of Blogs and similar Web portals [12, 16, 17, 18]. For the data from other social networks, such as MySpace, Facebook or Twitter monopartite networks are more suitable.
As we show in the following Sections, the networks obtained from the Blog and Digg

datasets are topologically inhomogeneous at the microscopic scale, at the level of node connectivity. Although, the bipartite network is sparse, its monopartite projection is dense and appear to have high clustering coefficient. The reason for this lies in the way of projecting certain motifs from bipartite networks onto monopartite ones. For instance, the star of size N, motif typical for sparse networks, is projected onto clique of the same size, type of motifs often present in highly clustered networks. 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 [40, 55, 51]. The methodology to systematically detect the communities in these types of networks based on the eigenvalue spectral analysis of the weighted bipartite graphs is described Section 2.1.2. These communities are important manifestation of the collective behavior on the Web.

3.2.2 Topology of networks and their projections

Degree of a node q, is given by the number of its first neighbors, the number of links between that node and the rest of the network [35]. In directed networks, one can distinguish between in- and out-degree of the nodes. In our bipartite networks, the node degree has a particular meaning for each partition. Specifically, for the user partition, the in-degree q_U of a node represents the number of posts and comments read (and commented), while the out-degree q_U out represents the number of posts and comments written by that user. Note that in the data the same user often comments a particular post (or comment) more than once, which is denoted by multiple link. Within the posts-and-comments partition, the out-degree q_P out of a node represents the number of users who left a comment on that particular post (comment), whereas the in-degree q_P in is the number of authors of the post (comment), which is always equal to one in the Blogs and Digg data.
As it was stressed, to describe structure of the network on the microscopic level one have to use two quantities, degree q and strength. In weighted bipartite networks, degree in user partition represents the number of different posts, commented by that user and its strength is equal to a number of comments posted on these posts. On the other hand, the degree of post represents the number of different users who commented that posts and strength is the number of its comments. In weighted monopartite networks of users/posts degree is the number of different users/posts connected to that node and strength is equal to sum of all links adjacent to that node.
Cumulative distributions for in- and out-degree for both partitions of the networks corresponding to the Blogs and Digg data are shown in Fig. 3.3. Broad degree distributions indicate large inhomogeneity both in the user- and post-partition. In particular, for the user partition, we have power-law dependences for both in- and

out-degree over two orders of magnitude, before the cut-off. Note that when the information about comment-on-comment is not available, that is BBC Blogs data, in- and out-degree distributions coincide, otherwise they are different. The observed exponent is compatible with the differential distributions of user degree as

$$P(q) \sim (q)^{-\tau} \exp(-(\frac{q}{q_0})^{\sigma}) \ , \tag{3.12}$$

with the exponent is found in the range $\tau \in [1.5, 2]$ and the cut-off q_0 and stretching σ are depending on the dataset. Similar behavior is obtained in user partition of Digg network with exponents ≈ 2.
The out-degree of a post is roughly equal to the number of users who wrote a comment (including comments-on-comments) of that post Fig. 3.3 (top right). There is difference in the slope of the decay of this distribution for the BBC Blogs data and B92 Blogs data. However, the qualitative features are similar: a power-law decay for less popular posts (with number of comments smaller than a characteristic value n^*_{com} (approximately $100 - 200$). In both cases a sharp bending of the distribution occurs for the popular posts, on which the number of comments exceeds n^*_{com}. The slopes of the power-law distributions are not universal. The described bending of the distribution indicates existence of two type of posts in the network: the *normal posts* whit number of comments less than n^*_{com} and *popular posts* which have more than n^*_{com} comments. As it was shown if Refs. [12, 16, 17, 18] the dynamics of collective behavior on these two groups of posts is different. This difference is reflected in community structure of networks and features of different time series, which will be shown in the following Sections. On the other hand, the cumulative distribution of out-degree of Digg stories, Fig. 3.3 (bottom right), shows different behavior. It exhibits q-exponential behavior described by function

$$P(q^P) = A(1 + \frac{q^P}{\sigma})^{-\theta} \ , \tag{3.13}$$

for degrees larger than 27, whit exponents $\theta = 10.17(5)$ and $\sigma = 995(6)$. The q-exponential function behaves as power-law for $\frac{q^P}{\sigma} >> 1$, indicating existence of the hubs in the network. Although, the out-degree distribution does not exhibit bending around certain value n^*, we will retain the division of stories, according to number of comments, on normal and popular posts.

Distribution of commons is another measure emanating from the bipartite representation of the network, and determines weight of links in a suitable monopartite projection. One can consider projections in both partitions [12, 16, 17, 18, 18]. In this thesis we are interested in the behavior of users in the enlarged Blog space. Therefore we will consider the projection on the user partition, where the commons, C^P_{ij} are defined as the number of common posts and comments per pair of users. Distribution shown in Fig. 3.4 of commons C^P_{ij} per pair of users obtained from the

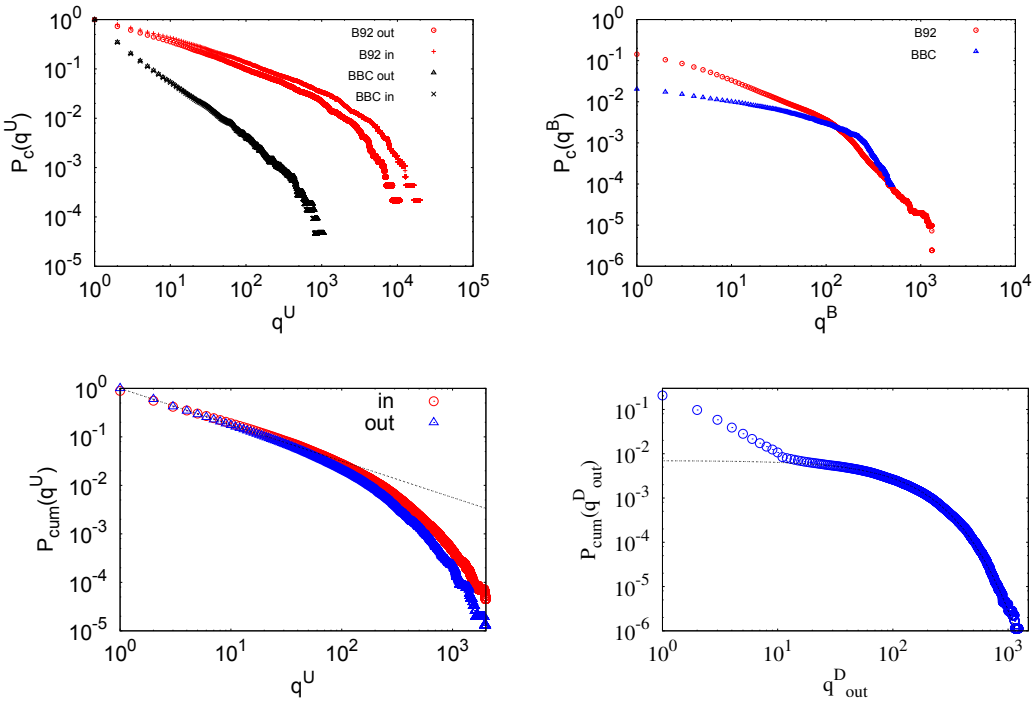

Figure 3.3: Cumulative degree distributions for users and posts from the bipartite networks representing data from (top) Blogs and (middle) Digg data. Cumulative distributions of degree (bottom left).

whole dataset, non-popular, and popular posts, as defined above, for B92 Blogs. The distribution has characteristic power-law decay for all sets of data. Similar features are found in the BBC Blogs data, but with an order of magnitude smaller cut-off. This also indicates that considerably larger overlap between users occurs on B92 Blogs compared to BBC Blogs, which can be attributed to much larger freedom in the authorship and the subject categories on B92 Blogs. The power-law distribution of commons indicates that large number of users is connected with links of small weight. By cutting links of small weight one can decrease network without loosing its important features such as community structure.

Mixing pattern (or disassortativity) can be calculated using one of the Eqs. 2.2, 2.4 and 2.5 for both node degree and strength depending on the data set we consider and network type. In directed bipartite networks one can distinguish between several combinations of the node's degrees which have particular meaning. For instance, we can consider in-out, in-in or out-out degree correlations. Out degree of of user node is given by the number of posts and comments written by user, while out degree of post gives the number of users who left a comment to that post. Out-out

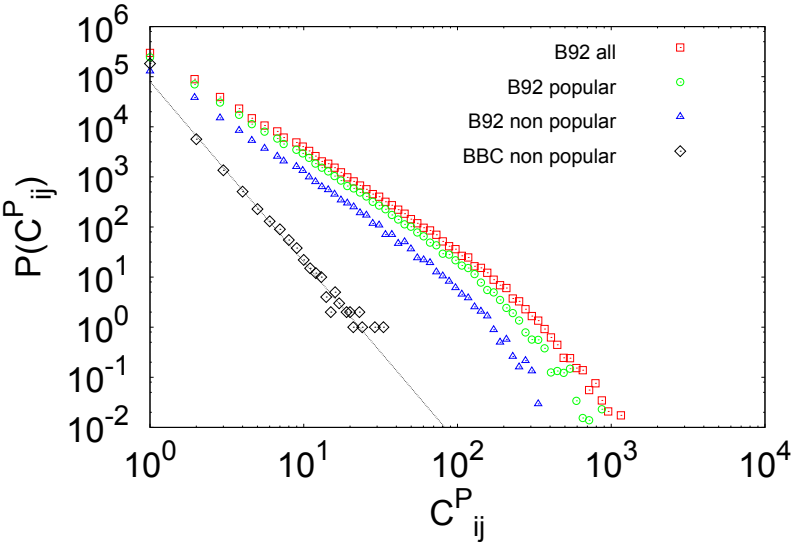

Figure 3.4: Distribution of commons: Number of common posts C_{ij}^P per pair of users for B92 and BBC Blogs.

degree correlations show weather very active users are active on very popular or posts with normal popularity.

To describe degree correlations for weighted networks one can consider three different measures (see Section 2.1.1). The Fig. 3.5 (right) shows mixing patterns for degree and strength correlations for weighted monopartite network of users related to popular BBC posts, i.e. posts with more than 100 comments. We find $N_P = 248$ popular posts and $N_U = 13674$ users who wrote $N_C = 53606$ comments on these posts. First we map this data onto directed bipartite network and then we find the projection of this network to a user partition by calculating matrix of commons C_{ij}^P. The average first neighbor degree and strength, calculated by Eqs. 2.2 and 2.4, are shown in Fig. 3.5. Both quantities have the same behavior as functions of degree/strength indicating assortative mixing in the network for the users with degree/strength smaller than 1000, i.e., users of similar activity tend to write comments the posts of same type. This type of mixing pattern suggests that users behavior in on-line social networks is similar to one in real life.

In bipartite networks mixing patterns show different behavior, more characteristic for technological networks. On Digg site users can leave comments-on-comment, which enables us to partition stories into two groups which have different dynamics: *discussion driven stories* which have more than 50% of comments-on-comments and stories driven by some external event. We select popular Digg discussions driven stories and users who commented and posted them, and create bipartite weighted

network of size $N_P + N_U = 3984 + 82201$. The average weighted first neighbor degree (see Eq. 2.5) for both user and stories partition decrease for small and high value of degree, Fig. 3.5 (right), indicating disassortative mixing in bipartite network for these group of nodes. The disassortative pattern suggest that users of small activity tend to leave comments on very popular stories, while very active users comment also posts with small activity which decrease weighted average degree. Similar patterns were found in other bipartite techno-social networks such as movie network [84]. Nodes with degree between 10 and 100 for users and 70 and 300 for posts are not correlated.

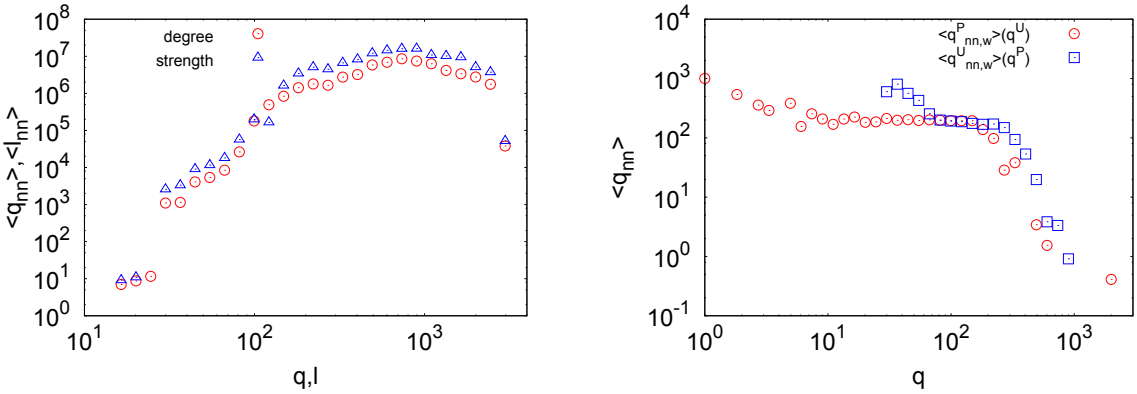

Figure 3.5: Dgree and strength correlations in weighted monopartite network of users on popular BBC posts (left) and weighted bipartite network of popular Digg discussions (left).

3.2.3 Community structure of techno-social networks

The described networks have one striking topological property, on the mesoscopic level they exhibit inhomogenities which occur as subgraphs of densely connected nodes [12, 16, 17, 18]. These communities play an important role in the collective dynamics in cyberspace.
Having the networks constructed from the data and put into a suitable form, as discussed above, we look for the community structure of these networks using eigenvalue spectral analysis method defined in Section 2.1.2. Specifically, for Blogs and Digg datasets we consider two types of representing networks:

- weighted projections of bipartite directed network onto partition of users or posts. Here network is represented with the matrix of commons C_{ij}.

- weighted bipartite network of posts and users, which is represented with matrix of weights W_{ij}.

The Normalized Laplacian for these networks is defined in a similar way as for binary networks, Eq. 2.12. The only difference, is that instead of adjacency matrix the suitable matrix representation of network is used, i.e.

$$L_{ij} = \delta_{ij} - \frac{C_{ij}}{\sqrt{l_i l_j}} \ . \tag{3.14}$$

The human social dynamics on *normal* posts differs form one on *popular*. As it will be shown, users on normal posts are grouped according to subject preference while on popular posts emotions have a leading role. For this reasons we will separately consider networks related to these two groups of posts.

Community structure of networks of normal posts

In BBC Blog the post has predetermined category for which reasons the community structure in users network is expected. We selected the group of normal posts with number of comments between 50 and 100. This gives $N_P = 149$ posts and $N_U = 4957$ users. We analyzed the Laplacian matrix for the respective projection on the user partition of the BBC Blog, given by Eq. 3.14. The resulting spectrum and scatter-plot are shown in Fig. 3.6. The scatter-plot exhibits three well separated groups of users, represented by large branches, and the central ring. Matching the identity of users at tips of these three branches, we then identify the lists of posts which they commented. These three users groups appear to be related to three different subjects of posts (*Sports*, *Business&economy*, and *Technology* Blog). Forth group is rather small and it is also related to *Sports* posts. In the case of B92 Blog no predefined category of post exists, and a community structure may appear in a self-organized manner. We consider a group of posts with less than 100 comments, consisting of $N_P = 3318$ posts. These posts together with corresponding comments constitute subgraph of size $N_B = 137941$, commented by $N_U = 3367$ users. Bipartite network constructed from non-popular group is then projected in the way described in the previous section and the monopartite weighted network of N_U users is determined and its spectra analyzed. Ranking of the eigenvalues of the respective Laplacian matrix and the scatter-plot in the space of three representative eigenvectors is shown in Fig. 3.7. Four groups (user communities) are clearly differentiated in the spectrum, and are marked as g_{U1}, g_{U2}, g_{U3}, g_{U4}. In the following we analyze the structure of Blogs to which these four user groups are related.
We inspect the text of the posts and their comments, for which the users in groups g_{U1} and g_{U4} (lower and upper branch in Fig. 3.7) are linked. This inspection reveals

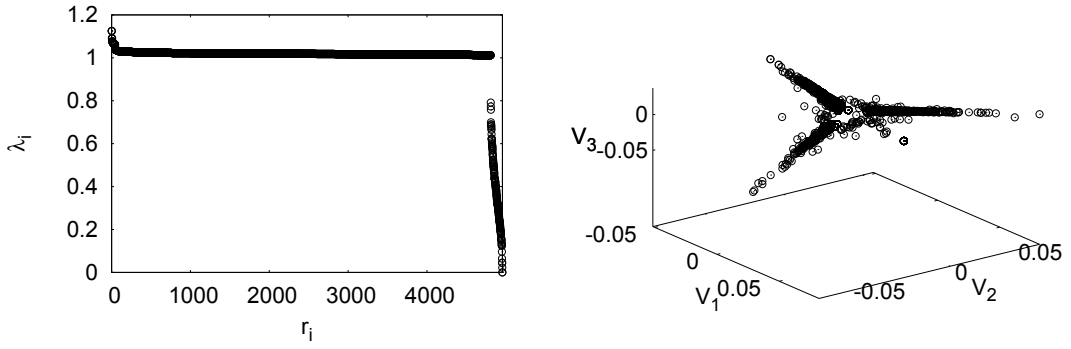

Figure 3.6: Spectrum (left) and the scatter plot (right) of three eigenvectors of the Laplacian for weighted user projected network of small nonzero eigenvalues, exhibiting three user communities.

that all posts in the user group g_{U1} are related to sports (football), and posts in group g_{U4} are about the urban architecture and related urban life problems [12]. The posts and comments related to the other two user groups g_{U2} and g_{U3}, are on mixed subjects and involve a larger community of users, however, they share more similarities apart from excluding the sports and the architecture subjects. The difference in posts related to both ends of these two branches is the time when they appeared: that posts related to the users on the left end of the scatter plot (g_{U3}) are posted at the beginning of the Blog site, as opposed to the right branch (g_{U2}), which are related to the recent posts.

Structure of communities on popular Blogs

In the case of popular B92 posts we consider weighted bipartite network. First, we filter the data according to the number of comments that exceeds 100 on a post and users who wrote these posts and the related comments. We then construct a bipartite network consisting of *users* and *posts*, while the comments are used to define the weights of links between them. The weight W_{ij} of the link between the user i and post j is defined by the number of comments that user i left on the post j. The constructed network for B92 blog consists of $N = 5079$ nodes (1466 posts and 3613 users). The weighted Laplacian matrix of the whole network is constructed as in Eq. 3.14 but with the W_{ij} weights and its spectrum is computed. The structure of communities in this network is shown in the $3-dimensional$ scatter plot of the eigenvectors in Fig. 3.8. In this projection four branches of nodes can be differentiated, denoted as $G_i, i = 1, 2, 3, 4$. Note that each point in this scatter plot

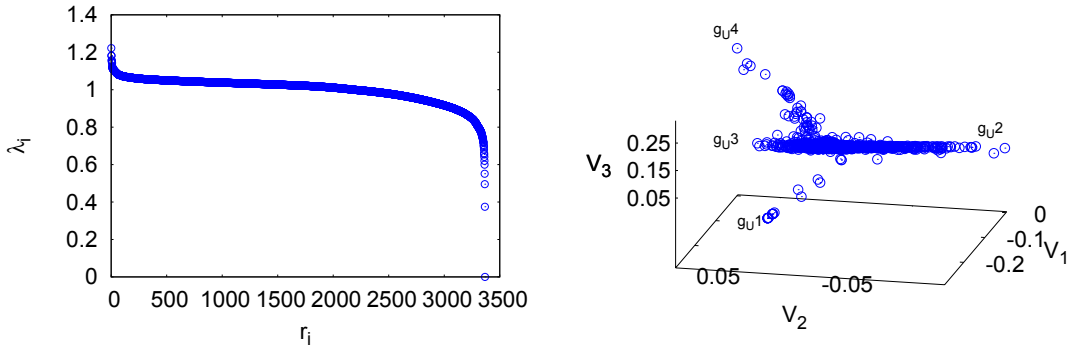

Figure 3.7: Eigenvalues in ranking order (left) and the eigenvectors of three lowest non-zero eigenvalues (right) of the weighted Laplacian matrix of the network of users on B92 Blogs. Four groups of users are marked

is either a user or a post. Here we are interested in the contents of the popular Blogs. For this purpose we further break the observed groups according to the spectrally detectable communities. We demonstrate it on the example of the network made of the groups $G_1 + G_2$ from the Fig. 3.8 (left). The corresponding structure of the communities is shown in Figure 3.8 (right). By inspection of IDb and the text on these posts, we find that the group G_1 of the previous plot remains the same (it is related to the Montenegrian political issues), while the group G_2 splits into three groups. A small group appears in the vertical plane contains posts related to in internal politics. In the other two branches are the posts related to the rights of pregnant women in hospitals in Serbia (left branch) and other still mixed issues and many users (right branch).

The community structure on popular posts (stories) can be also investigated trough projected network of users [16, 18]. For the case of BBC Blog the bipartite network of users and popular posts, consisting of $N_U + N)P = 13674 + 248$ nodes is projected onto user partition by the described procedure. Since the weight of the link between two users, given by the number of their common posts, has power-law distribution, shown in Fig. 3.4, the network can be considerably reduced by cutting the links with small weight. Specifically, keeping the links with weight $C_{ij}^P > 1$, 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. Figure 3.9 shows the spectrum and 3-dimensional scatter plot for smallest non-zero eigenvectors. Among four groups shown in Fig. 3.9 (right), we find that the largest group g_4 consists of users who often comment different types of posts, while the remaining three groups

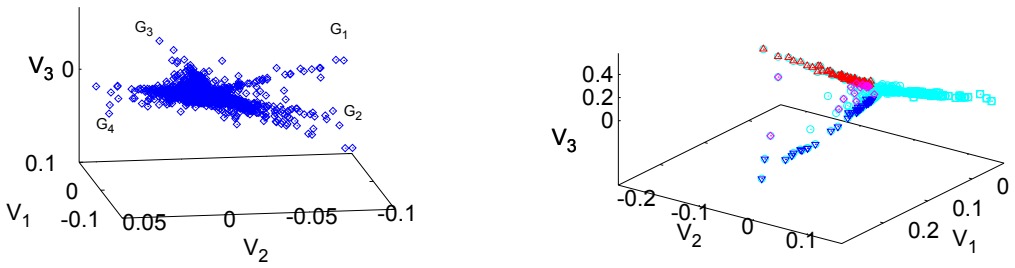

Figure 3.8: (left) Community structure in the bipartite network of users + popular posts from the data of B92 Blogs. (right) Further splitting in the structure of groups $G_1 + G_2$.

(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. Again, as in the case of B92 popular posts, we find the group of very active users. This type of group also appears in a network of users related to very popular movies [84], which suggesting another universal pattern of user behavior at popular posts.

In Fig. 2.1 (top right) three of the four user groups (g_1, g_2 and g_3), identified

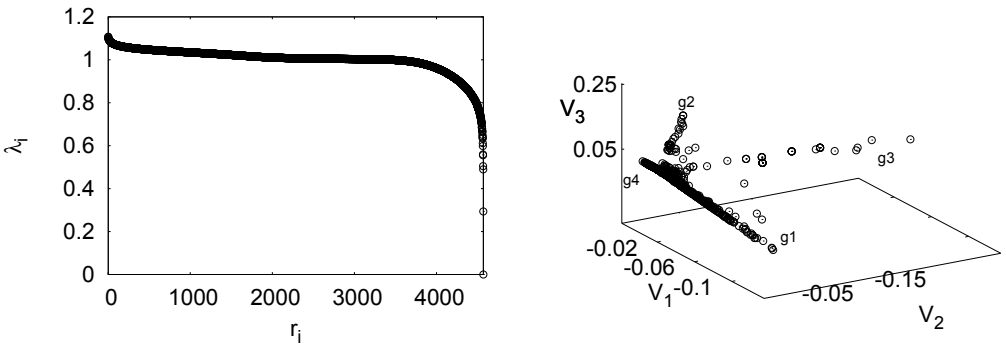

Figure 3.9: For the weighted user-projection of the bipartite network of the popular BBC posts: spectrum (left) and 3-dimensional scatter plot of the eigenvectors corresponding to three smallest non-zero eigenvalues (right), indicating four distinct user communities.

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 shown network consists of $N_U = 424$ users and $N_P = 50$ posts. User can write more than one comment on the post, and this comment can be positive (+1), negative (−1) or neutral (0). For every post we calculate the average emotional content according to $E_{i_P} = \frac{Q}{\sum_{j \in S_{i_P}} e_j}$, where e_j is emotional content of the comment j and S_{i_P} is a set of comments related to post i_P. The Q is the charge of emotional of comments related to post i_P, i.e., $Q = N^+_{i_P} - N^-_{i_P}$, where $N^\pm_{i_P}$ is the number of positive/negative comments. The color of a post represents its average emotional content, after the whole evolution time (within the dataset). A threshold for neutrality is applied for the average value $0.25 < E_{i_P}$. A few gray posts occurring in Figure 2.1 (top right) 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.

Unlike BBC Blog, Digg dataset contains information about weather the comment was related to original story or to one of its comments. This allows us the reduce down further the subgroup of popular stories and take into account only ones which have more than 50% of comments-on-comments. The reasons for this lies in the fact that we are mostly interested in collective behavior of users due to discussion dynamics. We will refer to this subset of popular stories as discussion driven Digg, ddDigg. The whole dataset consists of 4654 stories from which $N_P = 3984$ are discussion driven. Since the size of the user partition related to this stories is beyond the capabilities of the algorithm we reduced its size to $N_U = 4918$, by keeping only the users with strengths $l_i > 100$. In order to unravel what posts (and comments) keep a given community of users together, here we perform the spectral analysis of the weighted bipartite network, where the matrix elements W_{ij} represent the number of comments of a user i to post j. In this way we identify the list of user and post-nodes belonging to a community. We identify three communities, according to scatter plot inf Fig. 3.10(right): the largest community G_1 (right branch) consisting of $N_P = 2248$ stories and $N_U = 3423$ users, G_2 (upper left branch) with $N_P = 108$ stories and $N_U = 128$, and G_3 (bottom left branch) with $N = 948$ nodes out of which $N_P = 587$ are posts. For each identified community we then select from the original data all comments on the posts made by the users in that community, together with their time of appearance and the emotional contents. The categories and sub-categories of the Digg posts are also predefined. Closer inspection of the posts identified in different groups do not show any type of regularity reading the post subject. As we will show in the following Section, communities of users found in networks of popular stories are closely related to emotional content of comments posted by users.

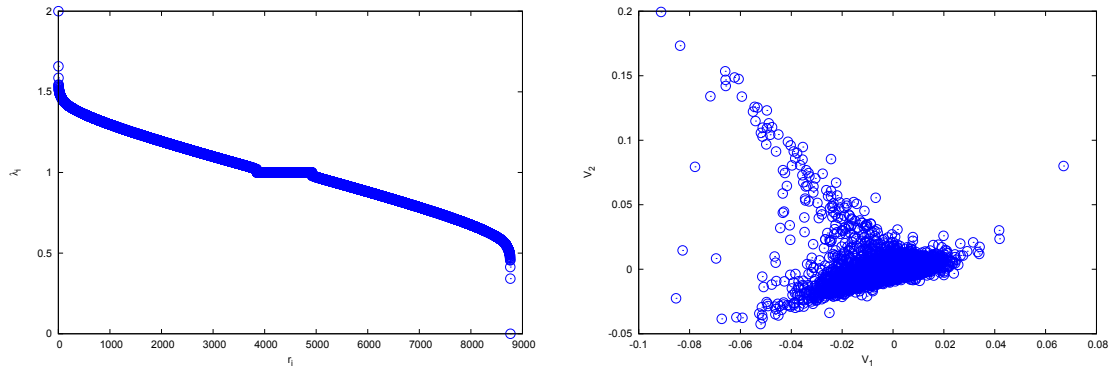

Figure 3.10: Spectrum (left) and scatter plot (right) of the eigenvectors corresponding to three lowest nonzero eigenvalues of the Laplacian indicating occurrence of user communities on ddDigg.

3.2.4 Emotional driven communities on popular posts

As it was mentioned, the dataset for B92 Blog does not contain information about emotional content of post and comments. For this reasons we are not able to fully investigate the dynamic of users clustering around popular posts. On the other hand, BBC Blog and Digg dataset contain this information which allows us to infer the role of emotions in formation of these communities and blogging dynamics.
It follows from distribution of number of positive, negative and neutral comments [16, 18], for BBC Blog and Digg datasets, that negative emotions are the most frequent, more than 50%, while there is less than 25% of positively classified texts. This indicates important role of negative emotions in Blogging dynamics. It is believed that the reason for high degree of negativity in communication on these Web sites is a consequence of the relative absence of social norms, or social control, which is characteristic for CM communication [115]. In CM communication one is less aware of the other person, basically because of the reduced visibility. In email exchanges for example, one generally knows ones interaction partners, but their presence is less salient. Often the lack of social norms is caused by the fact that one doesn't know the other person, which is the case in the majority of sites where human exchange their opinions, such as Blogs and Forums [115]. In both cases, when one does not know the other person and when one is also less aware of his or her presence, the anonymity of the situation is increased. For this reasons, one might get the impression that interacts with entities that do not have feelings and that therefore behavior, which is not desirable in face-to-face communication,

is normal.
We are interested in temporal evolution of communities found on popular BBC and Digg posts. Based on the emotional content of the comments related to each specific post, we define two variables, emotional charge $Q(t)$ and total number of emotional comments $Q_\nu(t)$, whereby we can measure and describe the emotional state of members in community over time. In particular,

$$Q(t) = N_+(t) - N_-(t) \quad , Q_\nu(t) = N_+(t) + N_-(t) \tag{3.15}$$

where $N_\pm(t)$ stand for the number of positive/negative comments on a given post at time t after its posting.
For every identified community we can extract comments posted by its members with information about posting times and emotional content. Based on this we can calculate the fluctuations of: number of active users, number of comments, charge and number of emotional comments. The Fig. 3.11 (left) shows the evolution of one community on popular BBC Blogs. 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. 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 users action. Similar behavior was found in the evolution of communities on popular Digg stories, Fig. 3.10. Again, the increase in the number of users is closely correlated with the excess of the negative comments (critique) on the posts Fig. 3.11 (right).

3.3 Temporal patterns

Beyond the topological features with the community structure, the origin and the evolution of the user communities is a question for which both the *patterns of user behavior* and the *emotional contents* of the posts and comments play an important role. In this section we address the question of temporal behavior of user and posts using methods described in Section 2.3.

3.3.1 Temporal patterns of User behavior

It was stressed that temporal patterns of user activity have a fractal structure, as shown for instance for the Blogs and Digg data in [12, 16, 17, 18]. This means that

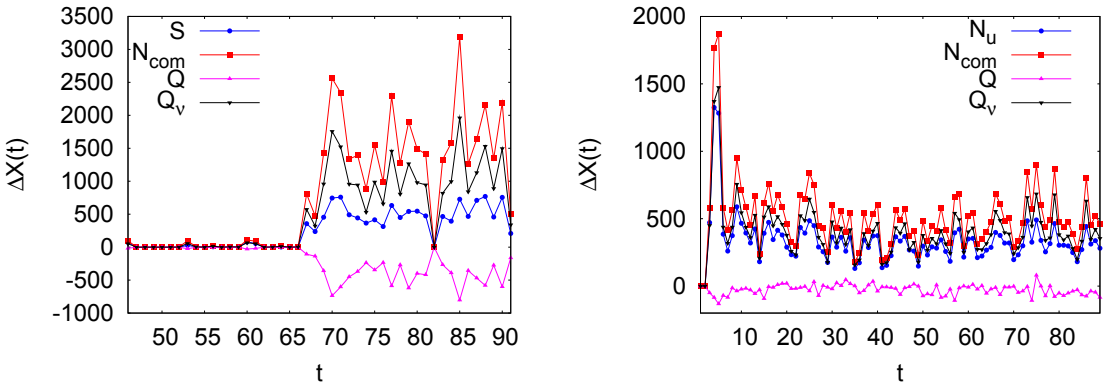

Figure 3.11: (left) For the large community from Fig. 3.9: fluctuations of the size of community plotted against time in weeks. Shown are also the fluctuations of number of all comments, number of emotional comments and charge of the posts. (right) Temporal fluctuations of the number of users, the number of comments, and the number and the charge of emotional comments of in the largest community (G_1) from Fig. 3.10. All variables are calculated for the time bin of one day.

the distance between two successive points in time Δt, that is user inactive time, obeys a power-law distribution

$$P(\Delta t) \sim (\Delta t)^{-\tau_\Delta} exp[-\frac{\Delta t}{\Delta t_0}] \tag{3.16}$$

whit exponent $\tau_\Delta \geq 1$ and cut-off depending on the size of dataset. We select the group of users which were active on BBC popular posts, and calculate the set of Δt for each user. Fig. 3.12 shows the distribution of Δt averaged over all users from the selected group. The distribution has power-law behavior, Eq. 3.16, with exponent $\tau_\Delta = 1.08(3)$ and exponential cutoff at $\Delta t_0 = 82310 \pm 5416$. Similar distributions were found for the other two datasets with the exponent $\tau_\Delta \approx 1.5$. For the same selected a group of users we also calculated the time intervals Δt between two successive positive (negative) comments made by a particular user at anyone of the posts. The distributions of these time intervals $P(\Delta t)$ averaged over all users are given Fig. 3.12 (right). The distributions again, exhibit power-law decay with stretched-exponential cut-offs,

$$P(\Delta t) \sim (\Delta t)^{-\tau_e} exp[-(\frac{\Delta t}{\Delta t_0})^\sigma] \ , \tag{3.17}$$

where the stretching exponent $\sigma = 0.8(1)$ and the slope $\tau_e = 1.01(3)$ for both positive and negative emotion comments.

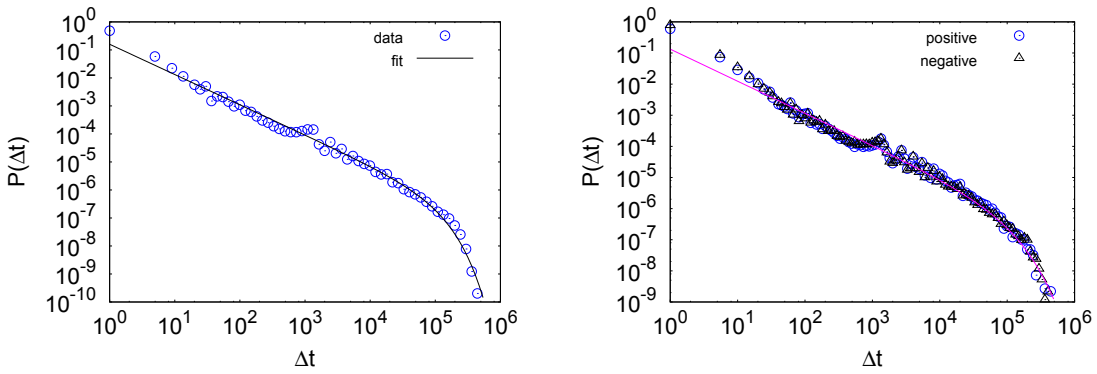

Figure 3.12: Distributions of time delay Δt between two consecutive user actions (left) and positive and negative actions (right), averaged over users related to popular BBC posts.

The fractal nature of user behavior can be visualized with activity pattern Fig. 3.13. The user indexes are ordered by the time of their first appearance in the dataset, hence the top boundary of the plot indicates the appearance of new users, relative to the beginning of the dataset. The profile of the top boundary shows that new users arrive in waves. Moreover, the arrivals of new users boost the activity of previous users, which is manifested in the increased density of points in depth of the plot below each wave. As in the case of B92 Blog Fig. 2.7 (right) newly registered users are active within some time intervals, whereas their activity is reduced in later times. Some users persist over long time period, while many other users either reduce frequency or stop writing on Blog or Digg altogether.

Further quantitative analysis on temporal behavior of users can be done by investigating the time series of number of comments of each users $n_{com}(i,t)$. An example of such time series of a very active user from B92 is shown in Fig. 3.14 (left), with time bin of one day. The power spectrum of the time series shows long-range correlations at large frequency region, which suggests that the user activity is correlated over small time intervals. One can also analyze these time series by calculating Taylor's scaling (see Section 2.4). In Fig. 3.14 (right) the dispersion of these time series $\sigma_U(i)$ of a given user i is plotted against its average over all time windows $< n^U_{com}(j) >$. Thus in this plot each point represents one user from our list. The plot obeys the scaling relation given in Eq. 2.13. It is interesting that all users in the considered Blogs on both BBC and B92 Blogs follow the same scale invariance, with the exponent $\mu \sim 0.88$. The observed scale invariance of the user time series in Fig. 3.14 strongly indicates non-randomness in the user activity on Blogs.

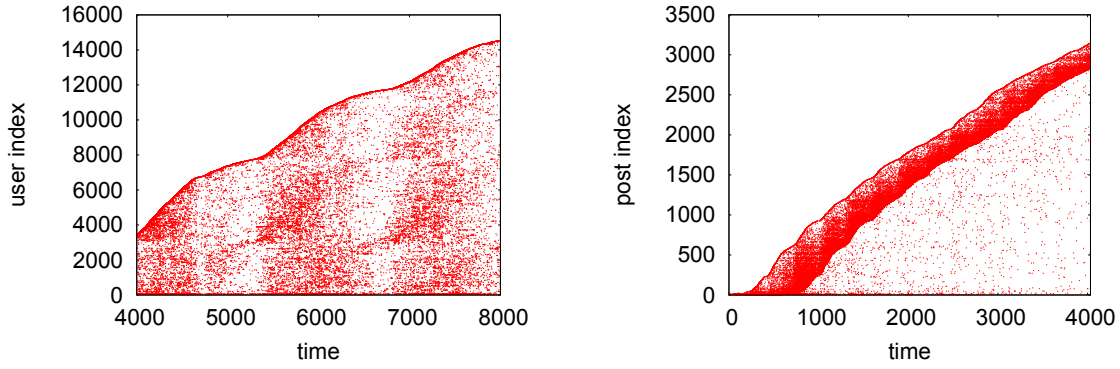

Figure 3.13: Example of temporal patterns of (left) user actions and (right) activity at posts, obtained from the original dataset of discussion-driven Digg (ddDiggs). Indexes are ordered by the user (post) first appearance in the dataset, while time is given in minutes.

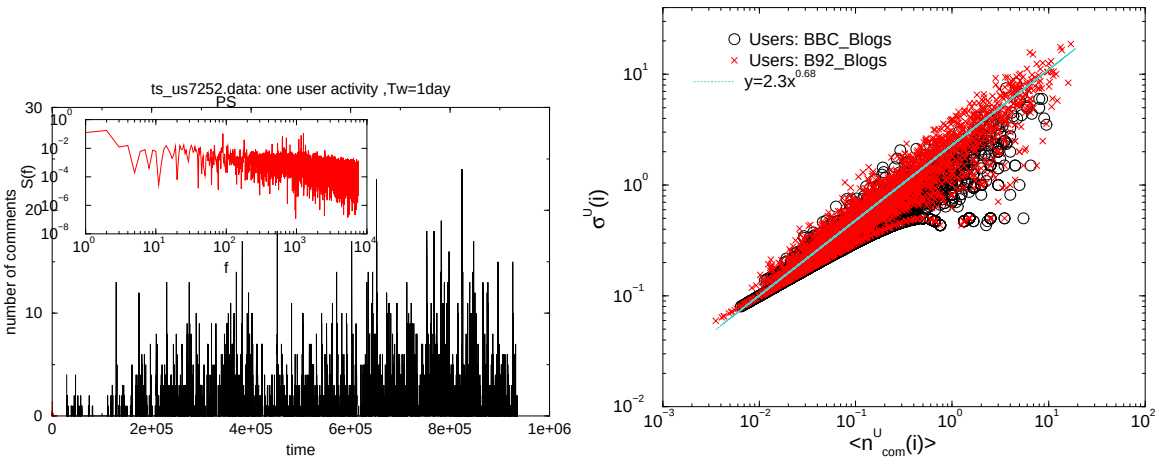

Figure 3.14: (left) Time series of an active user on B92 Blogs and (Inset:) its power spectrum. (right)Scaling of the dispersion $\sigma^U(i)$ of users activity (time bin one day) time series plotted against its average value $\langle n^U_{com}(i)\rangle$ for all considered users on B92 and on BBC Blogs.

3.3.2 Temporal patterns on Posts

Similar temporal patterns can be observed for each post. The linking pattern for ddDigg is shown in Fig. 3.13 (right) the points on the posts pattern indicate the times when an activity occurred at that post by anyone of the users. In the post-activity pattern dense points in a narrow time window following the post appearance time indicate an intensive activity at that post. This might be related with certain

exposure of the posts to users during that time period.
For the BBC and B92 Blog data we construct the distributions $P(t - t_i)$ where t_i denotes the moment when the post was published and t corresponds to a moments when someone wrote a comment on that post. The distributions averaged over all posts are shown in Fig. 3.15 (left) for both B92 and BBC Blogs, with slopes 2.8 and 2.3 respectively. Again, its remarkable that the power-law tail in both cases exists, which can be fitted with the q-exponential expression

$$P(t - t_i) = C(1 - (1 - q)\frac{t - t_i}{t^*})^{\frac{1}{1-q}} \ . \tag{3.18}$$

The differences in the exponents might be attributed to closing the posts in B92 after a preset expire date, which is not the case with the posts on BBC Blogs. In reference [95] an exponent larger than two is expected when the subject is repeatedly brought to peoples attention, which also might be the case with the highlighting recent (or very active) posts on the Blogs.
One can construct the time series of the number of comments (within a time bin) on a post. The time series of typical post in Fig. 3.15 (right) shows an increase shortly after the post appearance, then it drops to a steady value and eventually stops, an example with time bin of 1 h is shown in the inset to Fig. 3.15 (right). The scaling plot of all posts time series, similar to the one of users in Fig. 3.14 (right), is shown in Fig. 3.15 (right). In contrast to users, here one can see that posts belong to two distinct categories according to the popularity and, consequently, their dispersion of the time series are different. The active posts show large average number of comments per time bin and at the same time large dispersion of the time series according to Eq. 2.13 with the exponent $\mu = 0.68$. Whereas a large group of posts both in BBC and B92 Blogs exhibits a random variation in the time series, these posts are represented by the lower left part of the plot, where the exponent is $\mu = 0.5$. The behavior is statistically similar on both Blog sites, with the exception of few very active posts in the B92 Blogs (appearing close to the tip of the plot).

3.3.3 Temporal patterns of emotions on popular posts

The *emotional charge* Q and *number of emotional comments* Q_ν have different temporal behavior on different popular posts. For BBC Blog 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 Fig. 3.16 (left) with two examples of posts, representing these different evolution patterns. Specifically, these are the posts named College 5 ($MA1227829$) from the category

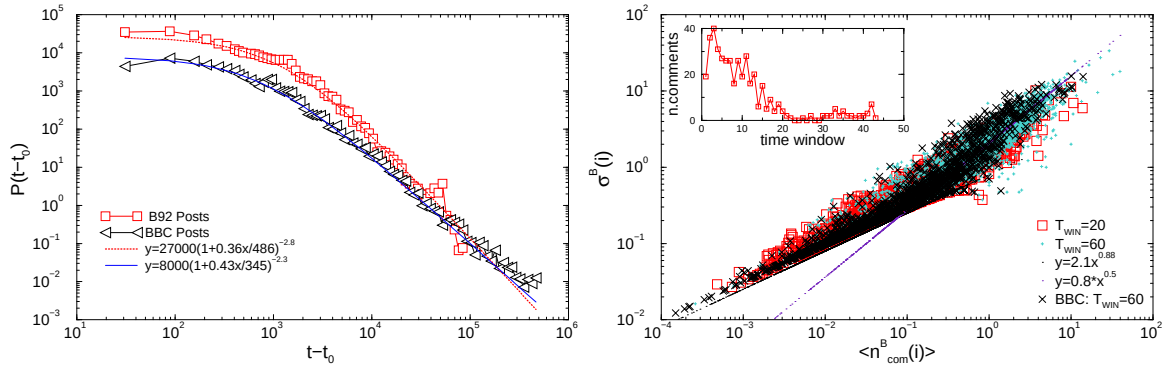

Figure 3.15: (left) Distribution of the response time to posts on B92 and BBC Blogs. (right) Dispersion vs average for posts with time bin 20 min and 1 h for B92 and for BBC Posts with 1 h bin.

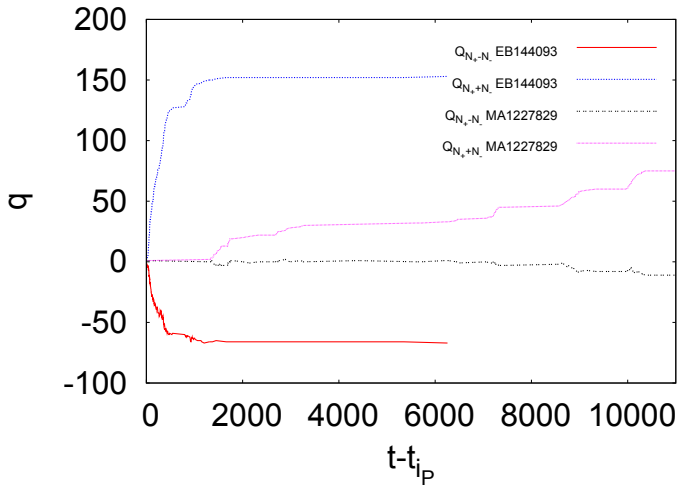

Figure 3.16: Evolution of the number of emotional comments Q_ν^B and charge Q^B plotted against time since posting, in two selected post with different patterns of popularity.

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 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 Fig. 3.17, 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.

Full information about users action over time and the contents of its comments

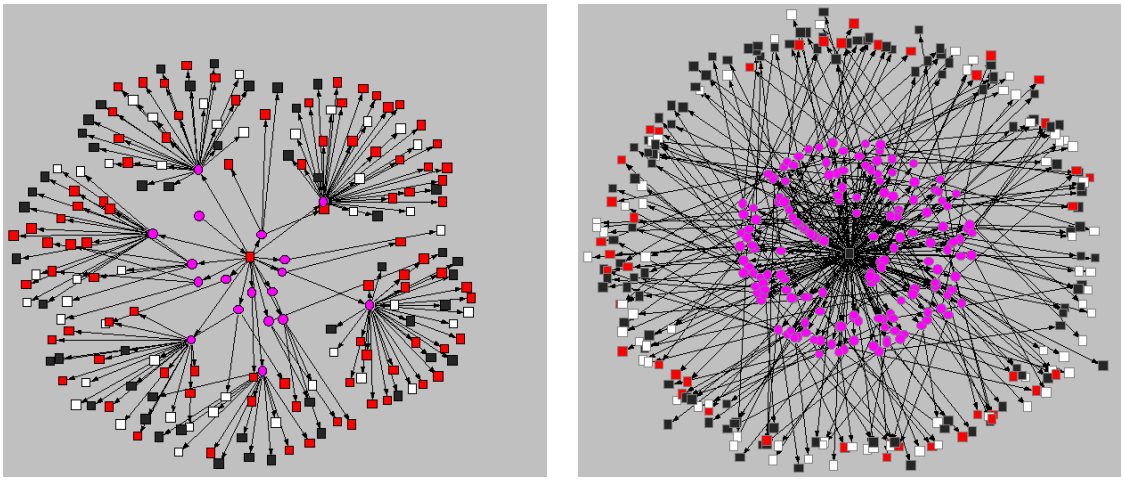

Figure 3.17: Two types of popularity: network of a discussion-driven popularity, music post with the title "College 5" (left), and of an externally-triggered popularity, the post is from the economy and business category titled as "Woolworth into administration" (right).

give an opportunity to analyze patterns of users behavior quantitatively. 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_ν^U, and the absolute values of negative charge of these comments, Q^U, as shown in Fig. 3.18. 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 Eq. 3.17 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.

Inspection of the time series of user activity with marked polarity shows that although the polarity of the comments by the same user often flips from positive to negative and back, the excess of the negative comments is found over the periods of users intensive activity (see for instance [16]). This explains the appearance of the

negative charge over larger time period, as shown in Fig. 3.18 for the same user.

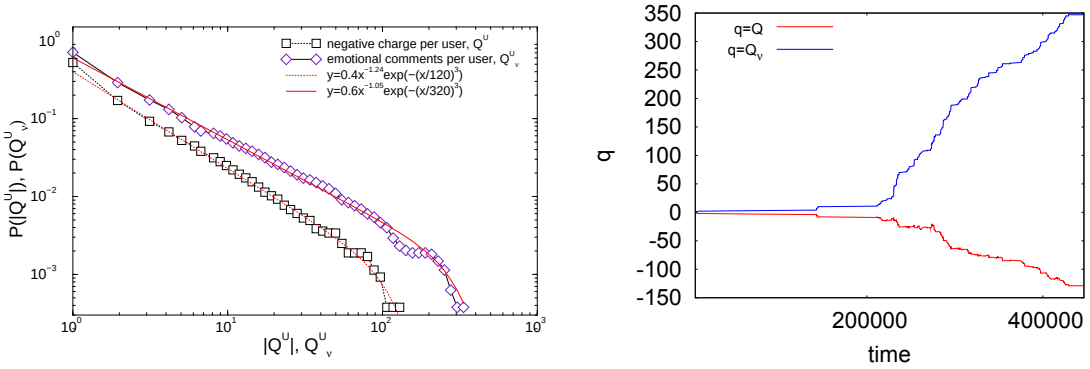

Figure 3.18: (left) Cumulative distributions of the number of emotional comments by a user, Q_ν^U, and of the negative charge of a user, Q^U, averaged over $N_U = 3000$ most active users in the dataset. The fit lines according to Eq. 3.17 are explained in the text. (right) The evolution of the number of emotional comments of any polarity, Q_ν^U, and their charge, Q^U.

3.4 Evidence of self-organized critical state

Dynamics in on-line community can be also described and studied trough the analysis of time series of user activity on all post, i.e., the number of comments within a small time bin is followed within the entire time period available in the dataset. Similarly, one can also determine in the similar way the number of emotional comments, $Q_\nu(t)$, and the number of negative/positive emotion comments, $n_\pm(t)$. An example of time series of number of emotional and negative comments in Digg is shown in Fig. 3.19 (left): zoom of the initial part of the time series is shown, indicating bursts (avalanches) in the number of comments. The occurrence of increased activity over a large period of time suggests possible formation of a user community around some posts. In this example, the intensive activity with avalanches of comments lasted over 2153 h, followed by reduced activity with sporadic events for another 1076 h.

Analysis of the time series reveals long-range correlations in the number of emotional comments over time. In particular, the power spectrum of the type $S() \sim \frac{1}{\nu}$ is found, both for the number of all emotional comments and the number of comments with negative emotion of the time series from Fig. 3.19 (left). The power-spectrum plots are shown in Fig. 3.19 (right). The pronounced peaks, which are superimposed onto the overall $\sim \frac{1}{\nu}$ law, correspond to daily and weakly periodicities of user activity.

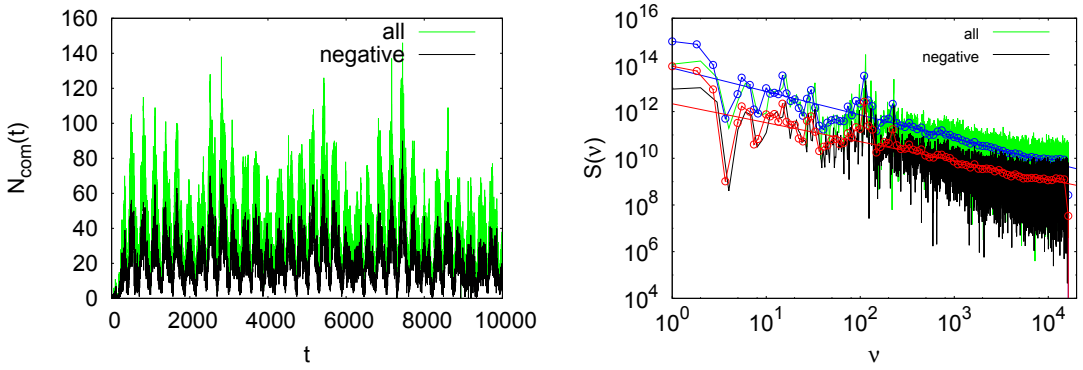

Figure 3.19: (left) Time series of the emotional comments-green, and the comments classified as carrying negative emotions-black, in the popular digg stories, plotted against UNIX time from the original data. (right) Power spectrum of the time series, with all emotional comments-green, and comments classified as negative-black.

The observed behavior of users is universal and do not depend on a specific site,

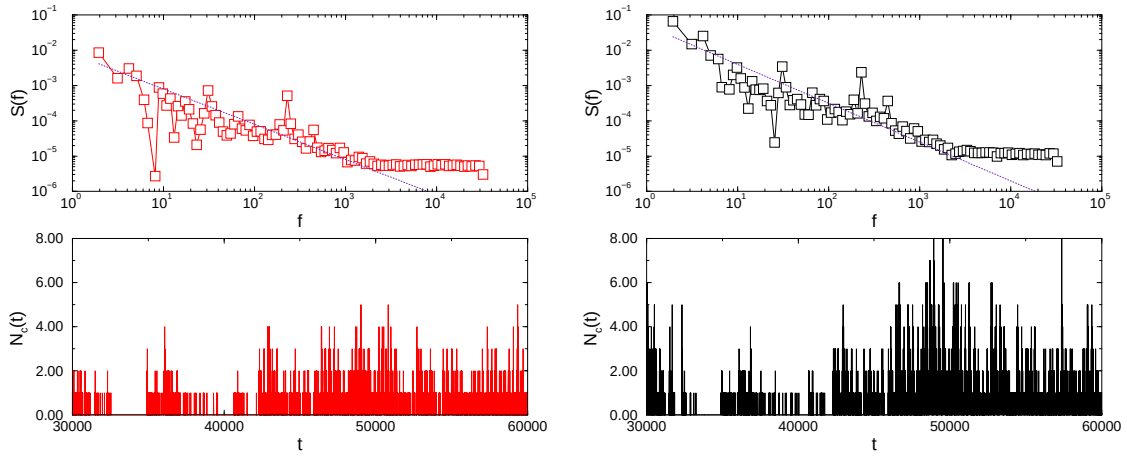

Figure 3.20: Example of the time series of the number of emotional comments $N_{\pm}(t)$ on the popular Blogs and their power spectrum $S(\nu)$: separated are comments which are classified as positive (left) and negative comments (right). Dotted lines indicate slopes -1.1 and -1.0, respectively.

Digg in this case. The time series of positive and negative comments on popular BBC posts are shown in Fig. 3.20. Again the power spectra of these time series appear to have $\frac{1}{\nu}$-type correlations beyond certain frequency, while the correlations vanish in the high-frequency range. Note that high-frequencies correspond to short time intervals. This indicates that users activities are correlated over long periods of times while for short times these correlations disappear. Apart from the abundance

of the negative comments, practically no differences can be detected in the spectrum of positive and negative comments.

The observed correlations in the time series are indicative of bursting events, which

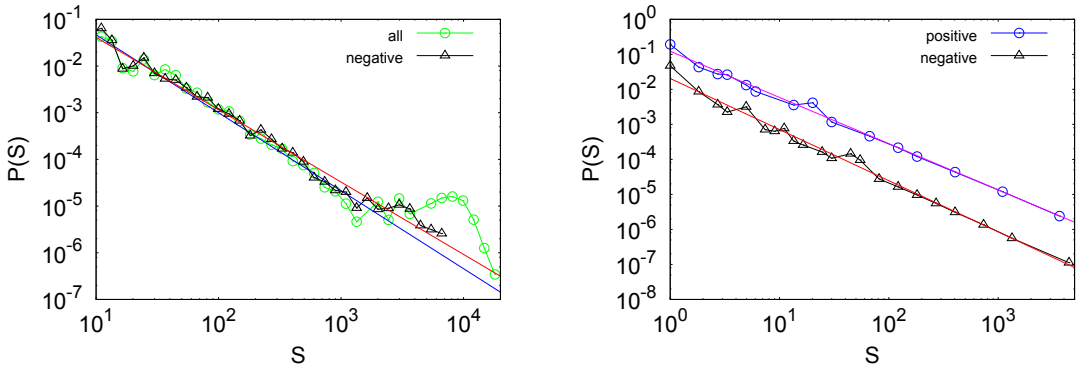

Figure 3.21: Distribution of avalanche sizes for time series of (left) number of all and negative comments (shown in Fig. 3.19) and (right) number of positive and negative comments on BBC popular posts (see Fig. 3.20).

are familiar to self-organized dynamical systems. As it was stresed, in our case an avalanche represents a sequence of comments, i.e., a comment triggering more comments within a small time bin t_{bin}, and so on, until the activity eventually stops. In analogy to complex systems as the earthquakes [102] or Barkhausen noise [104], the avalanches can be readily determined from the measured time-series, like the one shown in Fig. 3.19(top). Specifically, putting a baseline on the level of random noise, an avalanche encloses the connected portion of the signal above the baseline. Thus the *size* of an avalanche in our case is given by the number of comments enclosed between two consecutive intersections of the corresponding signal with the baseline. The distribution of sizes of such avalanches is shown in Fig. 3.22, determined from the signal of emotional comments from Fig. 3.19(left). A power-law with the slope $\tau_s \simeq 1.5$ is found over two decades.

The scale-invariance of avalanches is a signature of *self-organizing critical (SOC) states* [67, 65] in dynamical systems. Typically, a power-law distribution of the avalanche sizes

$$P(s) \sim s^{-\tau_s} exp(-s/s_0) \tag{3.19}$$

and other quantities pertinent to the dynamics [66, 116, 104] can be measured before a natural cut-off s_0, depending on the system size. The related measures, for instance the distribution of temporal distance between consecutive avalanches, $P(\delta T)$, also exhibits a power-law dependence, as found in the earthquake dynamics [102, 103].

Here we give evidence that the SOC states may occur in the events at individual posts in our dataset. We analyze avalanches on each post by bisecting post-by-post

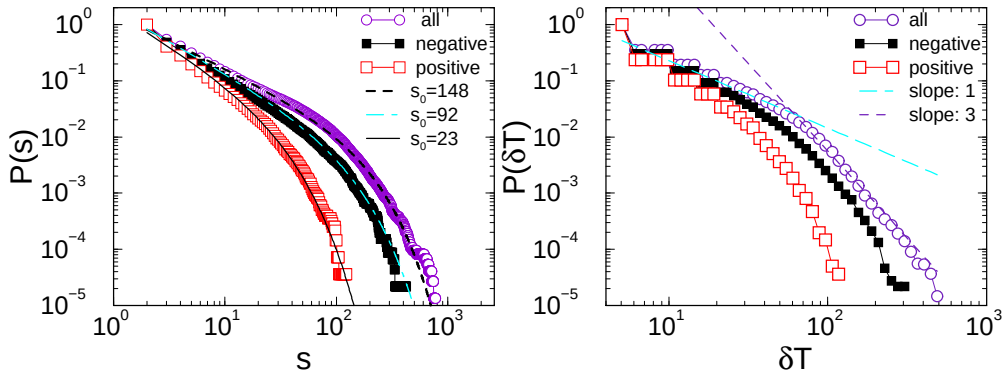

Figure 3.22: Cumulative distribution of avalanche size $P(S)$ (left), and inactivity time $P(\Delta T)$ between avalanches of comments observed at individual posts, bisected post-by-post from the network of ddDigg. The cases with comments of positive and negative emotional content are shown separately.

from the rest of the network. The results of the cumulative distributions of the avalanche sizes, $P(s)$, averaged over all 3984 posts, are shown in Fig. 3.22. The distributions for avalanches of different emotional contents are fitted with Eq. 3.19 with different exponents ($\tau_s \in [1.0, 1.2]$) and cutoffs. On the single-post networks we can also identify the quiescence times between consecutive avalanches: the distributions, $P(\delta T)$, are also shown in Fig. 3.22. It should be stressed that, besides the natural cutoff sizes, these avalanches are additionally truncated by the single post network sizes. Nevertheless, they can be fitted by the expression 3.19, indicating the self-organized dynamics at single-post level.
For comparison, the differential distribution of the avalanche sizes in Fig. 3.21 (left), which refers to the simultaneous activity on all posts, shows a power-law decay over an extended range of sizes, with exponents $\tau = 1.67(6)$ for all and $\tau = 1.46(5)$ negative comments, and an excessive number of very large avalanches (supercriticality). The activity on popular BBC posts is much smaller compared to one on ddDigg which results avalanches of small sizes. The distribution of avalanche sizes, Fig. 3.21 (right), exhibits power-law behavior with exponents $\tau_+ = 1.32(2)$ and $\tau_- = 1.46(3)$ for positive and negative avalanches respectively. Note that these data are logarithmically binned for better vision.

Chapter 4

Data driven models of User's Collective Behavior at Blogs

4.1 Modeling avalanche dynamics

In order to understand the origin of the critical states in the empirical data of ddDigg, and their dependence on the user behavior, we designed a cellular-automaton-type model on weighted post-user network, within which we simulate the dynamics, identify the realistic parameters which govern it and explore their effects by varying their values.
Emotional collective state has features typical for the systems exhibiting self-organized critical behavior [18] and cellular automata models can be used for studying the development of this state. Since the structure of social networks is irregular, to study social dynamics on these networks one needs to use modified type of CA model which is often refereed as *network automaton model*. In this model the cells are represented as nodes, while the neighborhood is formed by the vertices directly linked to a specific node.

The microscopic dynamics on Blogs, i.e., a user posting a comment, triggering more users for their actions, etc, can be formulated in terms of update rules and constraints, which affect the course of the process and thus the emergent global states. A minimal set of control parameters governing the dynamics with the emotional comments is described below and extracted from our empirical data of ddDigg. Specifically:

- User-delay Δt to posted material, extracted from the data is given by a power-law tailed distribution $P(\Delta t)$ in Fig. 4.1 (left), with the slope $\tau_\Delta = 0.89(5)$ above the threshold time $\Delta t_0 = 20603 \pm 1546$ min.

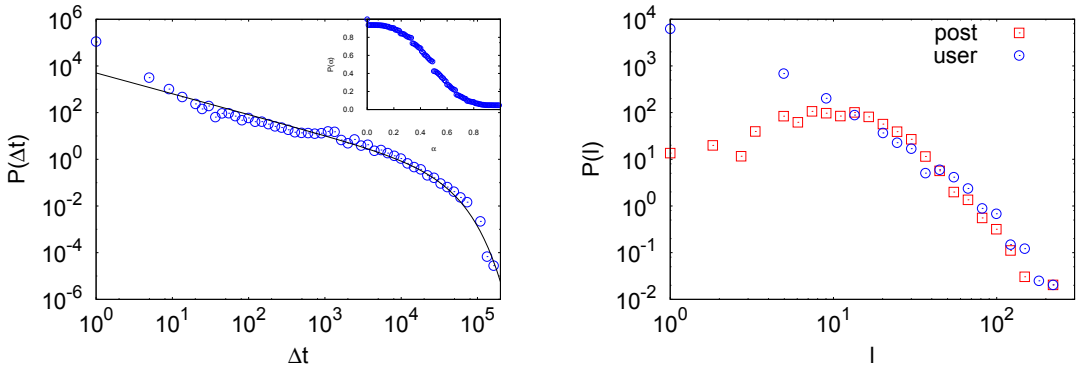

Figure 4.1: (left) Distribution of time delay Δt between two consecutive user actions, averaged over all users on ddDigg. Inset: Distribution of the probability α for a user to post a negative comment. (right) Distribution of strengths l_U,l_P for user and post-nodes on the weighted bipartite network, part of which is shown in Fig. 4.2.

- User tendency to post a negative comment, measured by the probability α, inferred from the data as a fraction of negative comments among all comments by a given user. Averaged over all users in the dataset, the distribution $P(\alpha)$ is given in the inset to Fig. 4.1 (left).
- Post strength l_P is a measure of attractiveness (relevance) of the posted material. Histograms of the strengths of posts and users in our dataset are given in Fig. 4.1 (right).
- User dissemination probability λ is a measure of contingency of bloggers activity. It is deduced from the empirical data as the average fraction of the users who are active more than once/at different posts within a small time bin ($tbin = 5$ min). In the model we vary this parameter, as explained below.
- Network structures mapped from the real data at various instances of time underlying the evolution of connected events. Here we use the weighted bipartite network, representing the G_1 community of the ddDigg data, (see Fig. 3.10). The network is visualized in the Fig. 4.2.

4.1.1 Model rules

Within the network-automaton model, whose pseudo code is given in Table 3, the parameters are implemented as follows: first, the weighted bipartite network is constructed from the selected data. To each post on that network we associate

Algorithm 3 Network Automaton Model: Program Flow

INPUT: Network W_{ij}; Parameter λ, T_{max} Distribution $P_c(\Delta t)$, $P(\alpha)$; Start Prompted-Users-List (PUL), Exposed-Posts-List (EPL);
Set initial conditions:
for all $1 \leq i \leq N_U$ **do**
  set user probability to negative comments α random from $P_c(\alpha)$ distribution
end for
for all $1 \leq i \leq N_P$ **do**
  set Post strength ℓ_P, exact form the empirical data
end for
select a User to place first comment on a Post connected to it and put that Post $\rightarrow EPL$;
for all $t < T_{max}$ **do**
  update PUL along the network links from EPL
  for all $i \in PUL$ **do**
    select Δt from distribution $P_c(\Delta t)$;
    if $\Delta t(i) > 0$ **then**
      user passive;
    else
      user active:
      select Post with $\ell_P > 0$ from EPL along network links
      (a) place negative comment on it with prob. α, else positive/neutral with equal prob.;
      (b) for each commented Post reduce strength $\ell_p - -$
      (c) each commented Post adds to EPL
      with probability λ select another Post along network links; repeat steps (a), (b), (c) on that Post;
    end if
  end for
  Sampling;
end for
END

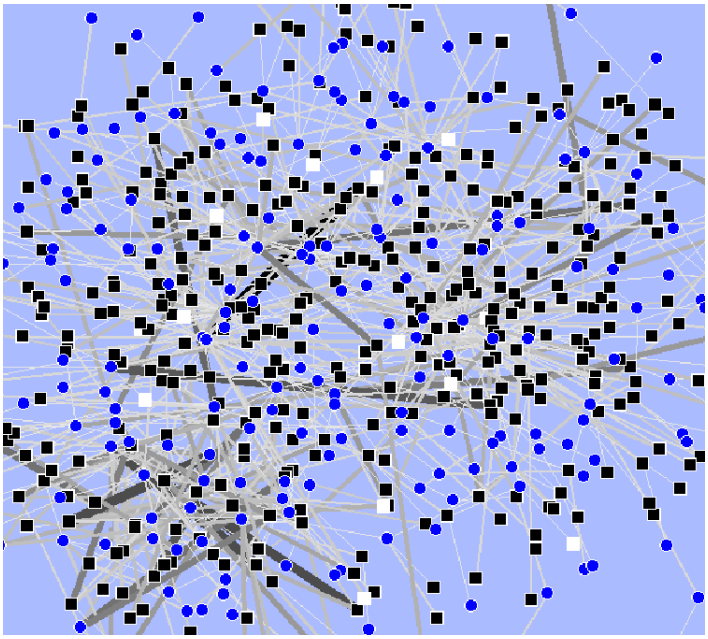

Figure 4.2: Part of a weighted bipartite network with users (bullets) and posts (squares). The widths of links are given by the number of comments of the user to the post. Color of the post-node indicates overall emotional content of all comments on that post.

its actual strength l_P, and to each user a (quenched) probability α taken from the actual distribution $P(\alpha)$. A well-connected user is selected to start the dynamics by posting a comment one of its linked posts. The lists of active users and exposed posts are initiated.

Then at each time step all users linked along the network to the currently exposed posts are prompted for action. A prompted user takes its delay time Δt from the distribution $P(\Delta t)$ of the actual dataset. Only the users who got $\Delta t < tbin$ are considered as active within this time step and may comment on one of the exposed posts along their network links. The posted comment is considered as negative with the probability α associated with that user, otherwise equal probability applies for the positive and objective comment. With the probability λ each active user may make an additional comment to any one of its linked (including unexposed) posts. The post strength is reduced by one with each received comment. Commented posts are added to the list of currently exposed posts. In the next time step the activity starts again from the updated list of exposed posts, and so on. Note that the activity can stop when: (a) no user is active, i.e. due to long delay time $\Delta t > tbin$; (b) the strength of the targeted post is exhausted; (c) no network links occur between currently active areas. Therefore, in contrast to simpler sandpile dynamics, in our

model the avalanche cutoffs may depend not only on the network size but also on certain features of the nodes and the links.

4.1.2 Simulation results: importance of dissemination

In the simulations presented here we vary the parameter λ while the rest of the parameters are kept at their values inferred from the considered dataset, as described above. The resulting avalanches of all comments and of the positive/negative comments are identified. The distributions of the avalanche size and duration are shown in Fig. 4.3 (left) for different values of the dissemination parameter λ.

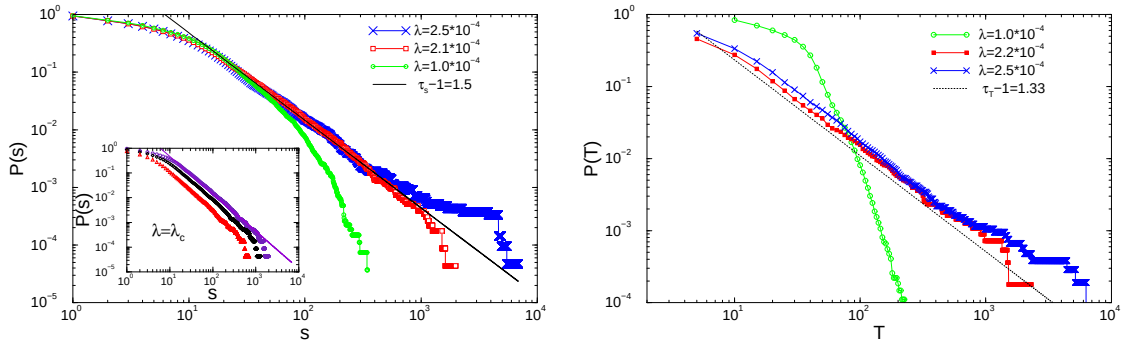

Figure 4.3: Cumulative distributions of the size (left) and duration (right) of avalanches of emotional comments simulated within the network-automaton model for different values of dissemination parameter λ. Fixed parameters for the network structure, post strengths and user inclination towards negative comments and delay actions are determined from the ddDigg dataset. Inset: the distribution $P(s)$ for all avalanches, and for avalanches of positive and negative comments for a critical value of the dissemination parameter $\lambda = \lambda_c$.

The simulation results, averaged over several initial points, show that the power-law distributions Eq. 3.19 of the avalanche size occur for the critical value of the dissemination, $\lambda = \lambda_c \sim 2.1 \times 10^4$, for this particular dataset, whereas varying the parameter λ in the simulations appears to have major effects on the bursting process. Specifically, the power law becomes dominated by the cutoff for $\lambda < \lambda_c$, indicating a subcritical behavior. Conversely, when $\lambda > \lambda_c$, we observe an excess of large avalanches, compatible with supercriticality. Therefore, our simulations support the conclusion that contagious behavior of some users may lead to supercritical avalanches, observed in the empirical data. The critical behavior at $\lambda = \lambda_c$ has been confirmed by several other measures. The slopes of the distributions of size and

duration, shown in Fig. 4.3 (right), are $\tau_s - 1 \approx 1.5$ and $\tau_T \approx 1.33$, respectively.

The critical behavior persists, but with changed scaling exponents, when the other control parameters are varied. For instance, we assume distribution of user delay to be

$$P(\Delta t) \approx exp(-\Delta t/T_0) \ , \tag{4.1}$$

and simulate avalanche dynamics for fixed $\lambda = \lambda_c$. Figure 4.4 shows cumulative distributions of avalanche sizes s and duration T for several values of parameter T_0. Evidently, both distribution exhibit the power-law behavior with the slopes τ_S and τ_T depending on the parameter T_0.

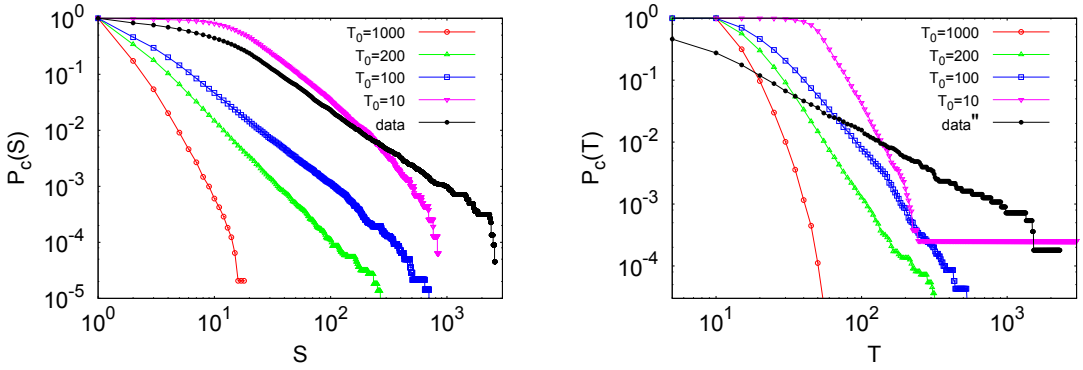

Figure 4.4: Cumulative distributions of the avalanche sizes, $P(s)$ (left), and avalanche duration, $P(T)$, (right) for exponential distribution of time delay given with Eq. 4.1, for several values of T_0. Filled symbols show the corresponding distributions obtained using the original distribution $P(\Delta t)$ native to the empirical dataset.

Beside distribution of user delays one can also vary distribution of parameter α with fixed values of all other parameters. It follows from the Fig. 4.5 that distribution of avalanche sizes is independent of the type of distribution for parameter α. As expected the parameter α influence only the number of negative/positive comments, and thus the size of maximal emotional avalanche. Distributions of avalanche sizes have power-law behavior with the same exponent as for $P(\alpha)$ obtained from the ddDigg dataset. In the case of exponential distribution given by

$$P(\alpha) = B \exp(-\alpha_0 \alpha) \tag{4.2}$$

with $\alpha_0 = 100$ the most probable value of parameter α are the one below 0.1, which results in small number of negative comments (see Fig. 4.5 (left)). Similar behavior is obtained for distribution

$$P(\alpha) = A\alpha^{-\tau_\alpha} \ , \tag{4.3}$$

with $\tau_\alpha = 2.1$ shown in Fig. 4.5 (right).

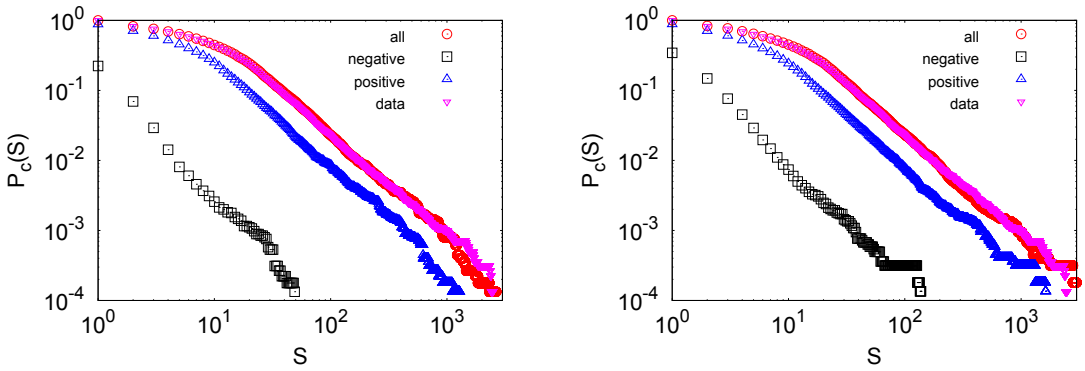

Figure 4.5: Cumulative distributions of the size of avalanches obtained by network-automaton model for fixed λ_c and $P(\Delta t)$ obtained from the ddDigg dataset, and varying the user inclination towards negative comments, $P(\alpha)$. (left) Exponential distribution given with Eq. 4.2 and $\alpha_0 = 100$ and (right) power-law distribution Eq. 4.3 and $\tau_\alpha = 2.1$.

4.2 Agent-based model of cuber-communities on bipartite networks

Agent-based modeling [117, 71, 4], where different properties of agents influence their actions, provides suitable theoretical framework for numerical simulations of social phenomena. Recently a model for product-review with the emotional agents in a mean-field environment has been introduced [118], with the agents emotional states described by two state variables (v_i, a_i). These variables correspond to the psychological values of the arousal and the valence, respectively, in view of the Russell's two dimensional circumplex model [112, 119, 120].
In the spirit of agent-based modeling, the agents (representing users on Blogs) are given certain properties that may affect their actions, i.e., the dynamic rules of the model. Conversely, these agent's properties are changed due to dynamic interactions between them, which imply further changed actions, and so on. To describe emotional actions of users on Blogs, we adapt the agents whose emotion states are described by two variables, arousal and valence, first introduced in Ref. [118]. Apart from the emotion variables, the agents in our model have additional properties indicated below in Eq. (4.4), and are subjected to the dynamically active networked environment, which affects their actions. Moreover, they are designed for multiple actions in course of the dynamics, thus contributing to the emergence of collective

emotional behavior.

In our model, the emotional agents are adapted to interact *indirectly* via posts on a bipartite network of the agents and posts. The essential elements of this model are:

- *2-dimensional local maps*, trough which we describe the dynamics of emotion variables, arousal and valence $\{a_i(t), v_i(t)\}$, of each agent;

- *Interaction environment*, represented by an *evolving bipartite network* of agents and posts, through which the agents emotion is spreading;

- *Driving noise*, applied to systematically perturb the system boosting its internal dynamics. In our model the system is driven by adding new users, according to the time-series $p(t)$ of new users, which is inferred from the empirical data of ddDigg, as explained above.

As it was stressed several times in this thesis, the users interactions on Blogs are not direct but trough the posts, for which reason the interaction environment is weighted bipartite network of agents and posts, where weight of the link between two nodes is equal to a number of comment of the user to the post. Unlike, network-automaton model (see Section 4.1) where the bipartite network is fixed and obtained from the ddDigg data set, in the agent based model network evolves due to the actions of/on agents. In this model, we actually have two types of agents, *user* and *post* agent, which can have different properties:

$$U[g; v_i(t), a_i(t); Lists_i..., \Delta t] \; ; \qquad P[t_P; < v_p(t) >, < a_p(t) >; Lists_p...] \; . \quad (4.4)$$

The dynamical variables arousal and valence $a_i(t)$ and $v_i(t)$ are properties of the user nodes, which vary in time as explained in detail below. In the moment of action, the agent transfers the values of its emotional variables (arousal and valence) to the post, thus contributing to the current overall emotional content on that post, $< v_p(t) >, < a_p(t) >$. Both user and post objects have *individual Lists* of connections on the evolving bipartite network, which are updated through the user actions, as explained below in Section 4.2.2. Additional properties that may affect the dynamics in our model are the post life-time, t_P, and the user action-delay, Δt, as well as the user probability for posting a new post, g. These properties, as in the case of network-automaton model, can be inferred from the considered dataset (for details see Section 4.2.3).
The dynamic rules of our model are motivated by the processes characteristic for

dynamics of Blogs and Digg, which are described in Section 3. Since we are interested in the discussion dynamics, which is result of the interaction between the agents and not of some external influence, the only driving force in the system is the number of new users given by $p(t)$. Unlike model given in Ref. [118], we do not take into account any additional external noise.
The interaction environment for our agents is defined with bipartite network, meaning that every agent has a unique list of connecting posts. The influence of environment on agent's emotional state is defined trough environmental fields in Eqs. 4.5-4.6 which are different for every agent. The network environment induces (and keeps track of) the heterogeneity among the agents in a natural way, through the lists of posts to which they were connected in the course of their actions.
First we will explain the dynamic rules of the model and all parameters that control dynamic, and then we'll show the simulation results for certain values of parameters.

4.2.1 Emotional states of individual agents

Following Ref. [118], we assume that the individual emotional state (arousal and valence) of each agent can be described by two nonlinear equations, which are subject of the environmental fields. For our system on bipartite networks, the arousal and the valence are associated with each user-node and their values, kept in the intervals $a_i(t) \in [0,1]$ and $v_i(t) \in [-1,1]$, are updated according to the following nonlinear maps:

$$a_i(t{+}1) = \begin{cases} (1-\gamma_a)a_i(t) + [h_i^a(t) + qh_{mf}^a(t)](d_1 + d_2(a_i(t) - a_i(t)^2))(1-a_i(t)) & \text{if } \Delta t_i < 1 \\ (1-\gamma_a)a_i(t) & \text{otherwise} \end{cases} \tag{4.5}$$

and

$$v_i(t{+}1) = \begin{cases} (1-\gamma_v)v_i(t) + [h_i^v(t) + qh_{mf}^v(t)](t)(c_1 + c_2(v_i(t) - v_i(t)^3))(1-|v_i|) & \text{if } \Delta t_i < 1 \\ (1-\gamma_v)v_i(t) & \text{otherwise} \end{cases} \tag{4.6}$$

where $i = 1, 2 \cdots N_U(t)$ indicates the index of user node and t is discrete time step which corresponds to some time bin.
The coefficients d_1, d_2 and c_1, c_2 characterize the maps themselves, while the network environment effects appear through two types of fields: the local fields $h_i^a(t)$ and $h_i^v(t)$, and the mean fields $h_{mf}^a(t)$ and $h_{mf}^v(t)$. Note that the local fields $h_i^a(t)$ and $h_i^v(t)$ vary not only in time but also from user to user, depending on their connections on the network, and due to evolution of the network itself. On the other hand, the values of the mean fields $h_{mf}^a(t)$ and $h_{mf}^v(t)$ also fluctuate in time, but are equal for all users in each time step. Both, local and mean fields, are calculated based on currently active posts. The current emotional atmosphere on the Blog site is described trough mean-fields, which contribute to overall fields with a fraction

$q \in [0, 1]$. The q is varied as free parameter in Eqs. 4.5 and 4.6.
A user-node receives the stochastic inputs from the network in certain instants of time, when the events occur in its network surrounding, and reacts to them with a delay. The *delay time* Δt of each user is counted continuously. When the delay time is smaller than the computational time bin $\Delta t < 1 \approx t_{bin}$, the user is *prompted for the update* of its arousal and valence according to the full expressions indicated by top lines in the Eqs. 4.5 and 4.6. Otherwise, the arousal and valence values are only relaxing, with the rates $\gamma^a = \gamma^v \equiv \gamma$.

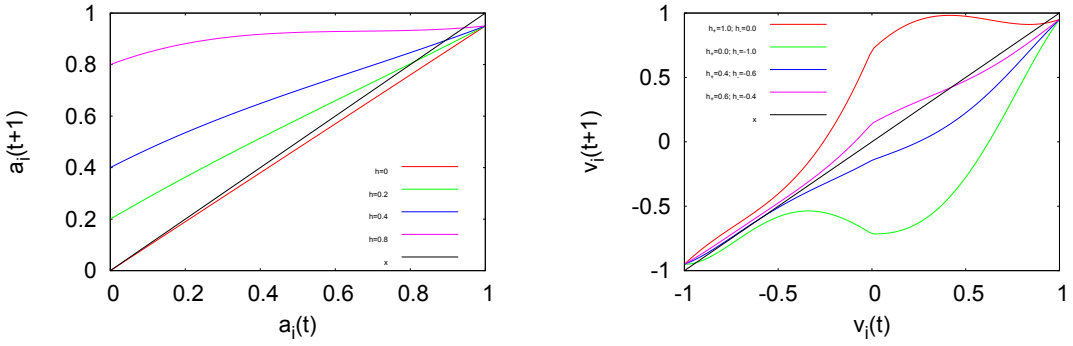

Figure 4.6: Maps for arousal (left) and valence (right) for four different values of the fields are shown. The fixed line $X(t+1) = X(t)$ is indicated.

For better understanding of dynamics at the local level, one can consider Eqs. 4.5 and 4.6 as two-dimensional nonlinear maps of an *isolated node* for a constant fixed value of the field. In Fig. 4.6 we show the situation for several values of the field pre-factors, while keeping all other parameters fixed to the values which are used later in the simulations. Depending on the parameters, the maps can reach different fixed points. In particular, the arousal map always leads to an attractive fixed point, the position of which depends on the strength of the field—larger arousal is reached when the field is stronger, cf. Fig. 4.6 (a). In the case of valence, two fixed points can be reached, one at the positive valence, when the fields are positive (upper branch), and the other one in the area of negative valence, which is attractive when the fields are negative (lower branches in the Figure 4.6 (b)). In general, when the nonlinear maps are coupled on a network, the network environment affects each individual map through a feedback loop, causing synchronization [53, 121] or other self-organization effects [122] among the nodes. In our case the network affects the fields, which thus fluctuate at every time step and depending on the node's particular position on that network. The dynamics of the fields can thus be visualized in the local map as jumping of the trajectories $v_i(t), a_i(t)$ from one branch of the map to another branch, and consequently, being attracted to another area of the phase

space. The nonlinear mapping takes part only when the agent is prompted to act, in the meantime, the maps are only relaxing towards the origin.

Agents interaction on network.

The fields $h_i^a(t)$ and $h_{mf}^a(t)$ in Eq. 4.5, which affect user i arousal at step $t+1$, are determined from the posts in the *currently active part of the network*, $\mathcal{C}(t,t-1)$, along the links of that user. Specifically,

$$h_i^a(t) = \frac{\sum_{p\in\mathcal{C}(t,t-1)} A_{ip} a_p^{\mathcal{C}}(t)(1+v_i(t)v_p^{\mathcal{C}}(t))}{\sum_{p\in\mathcal{C}(t,t-1)} A_{ip} n_p^{\mathcal{C}}(t)(1+v_i(t)v_p^{\mathcal{C}}(t))}; \quad h_{mf}^a(t) = \frac{\sum_{p\in\mathcal{C}(t,t-1)} a_p^{\mathcal{C}}(t)}{\sum_{p\in\mathcal{C}(t,t-1)} n_p^{\mathcal{C}}(t)} \ , \quad (4.7)$$

where $a_p^{\mathcal{C}}(t)$ and $v_p^{\mathcal{C}}(t)$ are the total arousal and the average valence of the post p calculated from the comments in two preceding time steps, while $n_p^{\mathcal{C}}(t)$ is the number of all comments posted on it during that time period. A_{ip} represents the matrix elements of the network, i.e., $A_{ip} > 0$ if user i is connected with the active post p, while $A_{ip} = 0$ if there is no link between them at the time when the fields are computed. Note that as network evolves trough time, the values of matrix elements A_{ip} can change, i.e., the link between post p and agent a may occur in later time as a consequence of agents action. In Eq. 4.7 the individual arousal fields $h_i^a(t)$ is modified by (dis)similarity in user's actual valence, $v_i(t)$, and the valence of recent comments on the post, $v_p^{\mathcal{C}}(t)$.

Regarding the valence fields in Eq. (4.6), we take into account contributions from the positive and the negative comments separately, while the neutral comments do not contribute to valence field. Depending on the current emotional state of the agent, positive and negative fields can lead to different effects [118], in particular, positive (negative) state will be influenced more with negative (positive) field, and vice versa. Here we assume that both components influence user valence, but with different strength according to the following expression:

$$h_i^v(t) = \frac{1-0.4r_i(t)}{1.4} \frac{\sum_{p\in\mathcal{C}(t,t-1)} A_{ip} N_p^+(t)}{\sum_{p\in\mathcal{C}(t,t-1)} A_{ip} N_p^{emo}(t)} - \frac{1+0.4r_i(t)}{1.4} \frac{\sum_{p\in\mathcal{C}(t,t-1)} A_{ip} N_p^-(t)}{\sum_{p\in\mathcal{C}(t,t-1)} A_{ip} N_p^{emo}(t)} \ , \quad (4.8)$$

where the valence polarity of the user i is given by $r_i(t) = \frac{v_i(t)}{|v_i(t)|}$, and $N_p^{\pm}(t)$ is the number of positive/negative comments written on post p in the period $[t-1,t]$. The normalization factor $N_p^{emo}(t)$ is defined as $N_p^{emo}(t) = N_p^+(t) + N_p^-(t)$. The mean-field contributions to the valence steam from the entire set of currently active posts $\mathcal{C}(t,t-1)$, and are independent on how users are linked to them:

$$h_{i,mf}^v(t) = \frac{1-0.4r_i(t)}{1.4} \frac{\sum_{p\in\mathcal{C}(t,t-1)} N_p^+(t)}{\sum_{p\in\mathcal{C}(t,t-1)} N_p^{emo}(t)} - \frac{1+0.4r_i(t)}{1.4} \frac{\sum_{p\in\mathcal{C}(t,t-1)} N_p^-(t)}{\sum_{p\in\mathcal{C}(t,t-1)} N_p^{emo}(t)} \ . \quad (4.9)$$

However, the mean-field effects are perceived individually by each user, depending on the polarity $r_i(t)$ of user's current valence.

4.2.2 Model rules

The rules of agents interactions on the network are formulated in view of user behavior on real Blogs and Digg and the observations from the quantitative analysis of the related empirical data shown in Chapter 3. In particular, the dynamic rules are motivated by the temporal patterns in Fig. 3.13 and the time-series in Fig. 3.19 (left), indicating how the number of active users arises in response to the arrival of new users. Moreover, additional features of the dynamics on ddDigg, suggest the dynamics with the dominance of negative emotions and with user's focus systematically shifting towards different posts.

In psychology literature human activity depends on their arousal, we assume the same in the model rules. We also use some of the general features of human dynamics, such as the occurrence of circadian cycles and delayed action to the events.
The pseudo code of the model is given in Table 4. The rules, implemented in $C++$, are given as follows. The system is initialized with typically 10 users who are connected to 10 posts, to start the lists of the *exposed* and the *active* posts and the *prompted* and the *active* users. Then at each time step:

- The system is driven by adding $p(t)$ new users (Note the correspondence of one simulation step with one $t_{bin} = 5$ min of real time); Their arousal and valence are given as uniform random values from $a_i \in [0, 1]$, $v_i \in [-1, +1]$, then updated with the actual mean-field terms. By the first appearance each user is given a quenched probability $g \in P(g)$ to start a new post. The new users are then moved to the *active user list*;

- The emotional states for all present users are relaxed with the rate γ, according to the second row in the Eqs. 4.5 and 4.6;

- The network area $\mathcal{C}(t, t-1)$ of the *active posts* is identified as post on which an activity occurred in two preceding time steps; we then update the emotional state of each agent i form the list of *prompted users* according to first row of Eqs. 4.5 and 4.6; prompted user is moved to a list of *active users* with probability $a_0 * a_i(t)$.

Algorithm 4 Agent Based model of emotional blogging:program flow

INPUT:time series $p(t)$; Distributions $P(\Delta t)$, $P(t_P)$, $P(g)$; n_t;
Set initial network of 10 agents and 10 posts connected randomly; initialize list of agents (LU) and posts (LP);
Initialize the list of prompted agents (LPA), list of active agents (LAA), list of exposed (LEP) and active (LAP) posts;
for all $t = 1, \ldots, nt$ **do**
  for all $k \in$ LU **do**
    Relax emotional state of agent k according to Eqs. 4.6 and 4.5;
  end for
  Calculate $h^v_{i,mf}(t)$ and $h^a_{mf}(t)$;
  for all $i = 1, \ldots, p(t)$ **do**
    Add new agent k_U and add it to NU; Initialize $List_{K_U}$;
    Chose initial $v_{k_U} \in [-1, 1]$ and $a_{k_U} \in [0, 1]$ and update Eqs. 4.6 and 4.5;
    Add Set its $\Delta T = 0$; g_{k_U} form $P(g)$; k_u to LAA;
  end for
  for all $k \in$ LPA **do**
    Calculate local fields h^v_i and h^a_i;
    Update emotional state of agent k according to Eqs. 4.6 and 4.5;
    With probability $a * a_0$ add user k to LAA;
  end for
  for all $k_U \in$ LAA **do**
    if q_{k_U} **then**
      Post j_P is created;$a_k(t) \to a^P_{j_P}(t)$ and $v_k(t) \to v^P_{j_P}(t)$;
      Initialise $List_{j_P}$; Add j_P to LAP;
    else
      Post a comment on post $j_p \in$ LEP (probability $P_p(t)$);
      The quantities $< v_{j_P}(t) >$ and $< a_{j_P}(t) >$ are updated;
      Update $List_{j_P}$ and $List_{k_U}$; Add j_P to LAP
    end if
    if μ **then**
      Post a comment on old post chosen according $p_{l_P,old}$;
      Update $< v_{l_P}(t) >$ and $< a_{l_P}(t) >$; Add l_P to LAP;
    end if
    for all $k_U \in$ UL **do**
      if $\Delta t_{k_U} > 0$ **then**
        $\Delta t_{k_U} - -$
      else
        Chose Δt_{k_U} from $P(\Delta t)$;
      end if
    end for
  end for
  for all agents k_U connected to a posts from LAP **do**
    if $\Delta t_{k_U} > 0$ **then**
      Chose Δt_{k_U} form $P(\Delta t)$;
    end if
  end for
  Update LEP; Update LAA;
end for
END

- Every active user:
  - adds a new post with the probability g or otherwise comment to one of the *exposed posts*, which are not older than T_0 steps; Users are linked to posts preferentially with the probability $p_p(t) = \frac{0.5(1+v_p^{\mathcal{C}}(t)v_i(t))+N_p^c(t)}{\sum_p[0.5(1+v_p^{\mathcal{C}}(t)v_i(t))+N_p^c(t)]}$, depending on the number of comments on it $N_p^c(t)$ and the valence similarity;
  - and with probability μ comments a post which is older than T_0 steps. The post is selected preferentially according to the negativity of the charge of all comments on it, with (properly normalized) probabilities $p_{j,old}(t) \sim 0.5 + |Q_j(t)|$, if the charge is negative, else $p_{j,old}(t) \sim 0.5$; Lifetimes of the posts are systematically monitored (already expired posts are not considered);
  - Current values of the valence and arousal of the user are transferred to the posted comment or the new post; User is given a new delay-time $\Delta t \in P(\Delta t)$; New posts are given life-time $t_P \in P(t_P)$.
- Delay-time Δt for all users that are not prompted is decreased by one and for prompted users we chose new Δt according to $P(\Delta t)$. Then, for all users that are connected to active posts (posted/commented in last two steps) we chose Δt once more if $\Delta t \neq 0$. The list of prompted users is then updated by adding all agents with $\Delta t = 0$.

4.2.3 Model parameters

According to the dynamic rules of the model, one can identify the parameters which control the dynamics at different levels. In particular, we use the following parameters, distributions or time-series which characterize, respectively:

- the local maps: $c_1 = d_1 = 1$, $c_2 = 2.0$, $d_2 = 0.5$, $\gamma = 0.05$;
- the properties of posts and users: $t_P \in P(t_P)$, $\Delta t \in P(\Delta t)$, $T_0 = 2days$, $\mu(T_0) = 0.05$, $g \in P(g)$;
- the driving: $\{p(t)\}$, $q = 0.4$, $a_0 = 0.5$.

We drive the dynamics by adding $p(t)$ new users in the system at every time step. The time series of number of new users, $p(t)$, per time bin ($t_{bin} = 5$ minutes) is calculated from the ddDigg dataset (see Fig. 4.7).

Beside $p(t)$ several other parameters and distributions can be also inferred from the high-resolution data of Blogs and Digg. Figure 4.8 shows the *distributions* of

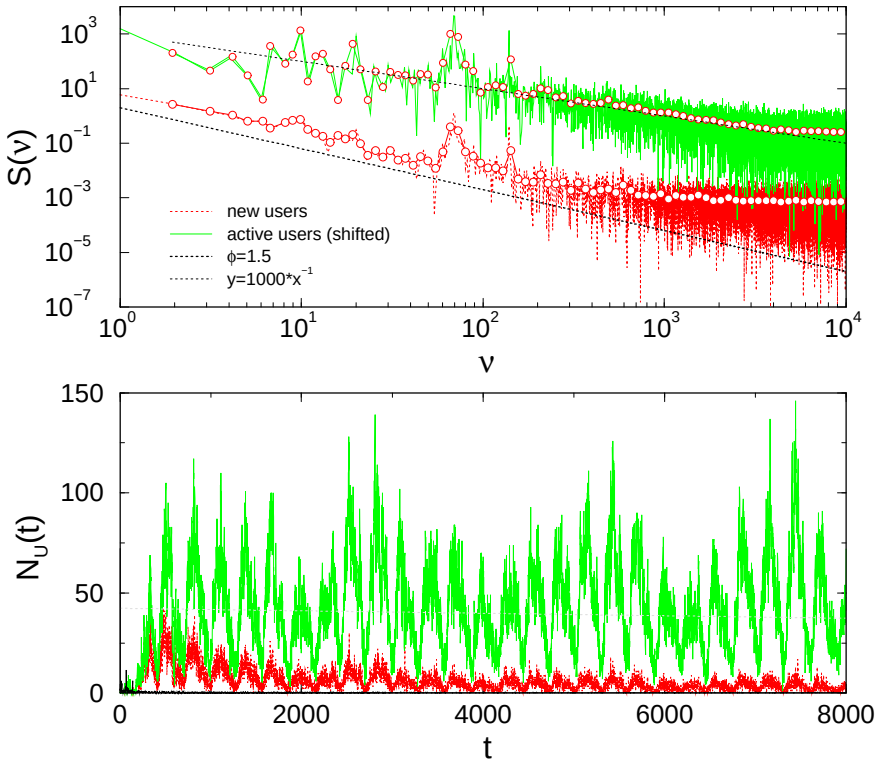

Figure 4.7: (bottom) Time-series of the number of new users (red-dark) and the number of all active users (green-pale) per time bin of $5min$ derived from the ddDigg dataset; (top) Power spectra of these time series as indicated (shifted vertically for better vision). Daily and weekly cycles can be easily noticed on both plots.

the life time of the post, $P(t_P)$, time delay of users actions, $P(\Delta t)$, and the fraction of new posts per user $P(g)$, as well as the *functional dependence* of dissemination probability $\mu(T_0)$ on time window T_0 which determines the subset of exposed posts. Parameters $P(t_P)$, $P(g)$ and $\mu(T_0)$ are obtained from Digg dataset, while distribution of time delays is derived from BBC Blog dataset. The value $T_0 = 576$ time bins (corresponding to 2 days of real time) can be approximately estimated from the posts activity pattern, cf. Fig. 3.13(right).
The numerical values of the remaining parameters can not be extracted from this kind of empirical data. Hence they are considered as free parameters, that can be varied within theoretical limits. The values quoted above are used for the simulations in this work.

Here we describe a general methodology how such parameters are determined from the empirical data of high resolution. The delay-time distribution $P(\Delta t)$ in Fig. 4.8c, is directly related with the user activity pattern, cf. Fig. 3.13 (left): for a given

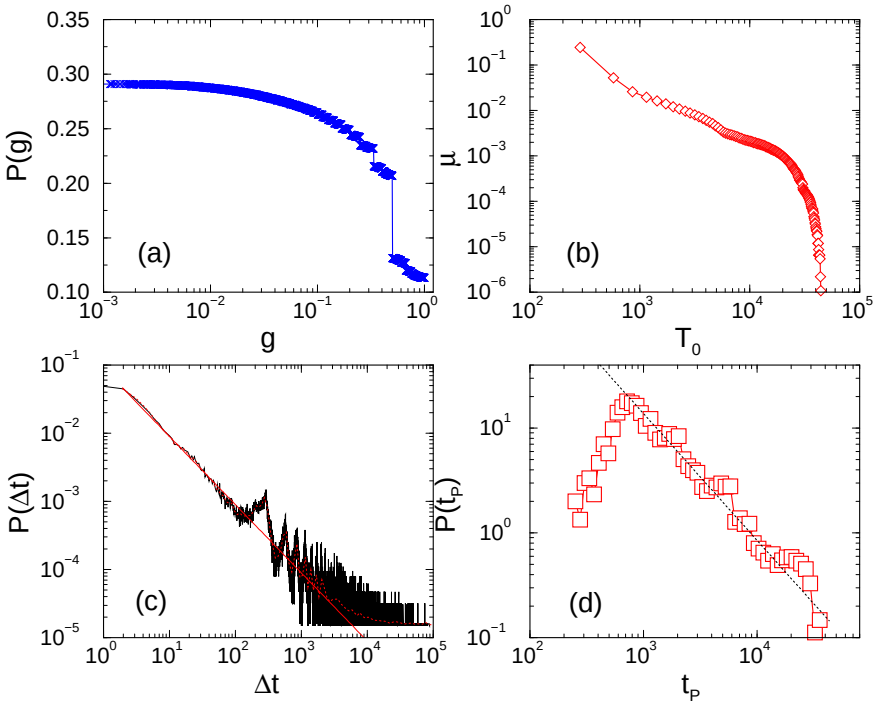

Figure 4.8: (a) Distribution of g—the fraction of new posts per user, relative to all posts on which that user was active, averaged over all users in the dataset. (b) Probability μ that a user looks at a post which is older than the specified time window T_0 time bins, averaged over all users and plotted against T_0. (c) Distribution of the time-delay Δt between two consecutive user actions, averaged over all users in the dataset. (d) Distribution of the life-time of posts t_P, averaged over all posts in the dataset (log-binned data). In Figs. (c) and (d) the time axis is given in the number of time bins, each time bin corresponds to $t_b = 5$ minutes of real time.

user (fixed index along y-axis) the delay time Δt is defined as the distance between two subsequent points along the time axis. The distribution is then averaged over all users in the dataset. Similarly, the distribution of the life-time of posts $P(t_P)$ in Fig.˙ 4.8 (d), is related to the pattern of posts activity as the distance between the first and the last point on the time axis for a given post, cf. Fig. 3.13 (right). The parameter T_0 is roughly estimated as the width of the time window, during which new posts were 'exposed' (dense points area in Fig. 3.13 (right)). When T_0 is fixed, then the probability that a user finds a post which is older than T_0 can be extracted from the data as the fraction of points beyond the dense area in the posts activity pattern until the post expires, cf. Fig. 3.13 (right). Then we have $\mu(T_0) = \frac{1}{N_P}\sum_{p=1}^{N_P}\left(\frac{1}{t_p}\sum_{t_{kp}>t_{0p}+T_0}^{t_p} 1\right)$, where N_P is the number of posts, t_p is the expire time of the post p, while t_{kp} and t_{0p} indicate the moments of the activity at

the post p and its creation time, respectively. Averaged over all posts in the dataset, gives the parameter $\mu(T_0)$, plotted in Fig. 4.8 (b) against T_0.
In the case of user properties, looking at the activity list of a given user, we can determine the fraction g of new posts that the user posted out of all posts on which the user were active in the entire dataset. The values appear to vary over time and users, the distribution $P(g)$ averaged over time and all users in the dataset is shown in Fig. 4.8 (a).
The values of the control parameters will depend on the empirical dataset considered. Specifically, the parameters as the life-time of posts, t_P, and users inclination to posting new posts, g, or to looking towards old material $\mu(T_0)$, strongly depend on the dataset. Note also that they might have hidden inter-dependences in view of the nonlinear process underlying the original dataset. For instance, if on a certain Blog site users are more inclined towards posting new material, which would yield increased probabilities of large g, then the life-time of posts may decline, resulting in a steeper distribution. Therefore, it is important to derive these parameters from the same dataset in order to ensure their mutual consistency.
On the other hand, as we showed in Chapter 3, certain features are universal and do not depend on the data set. For instance, the power-law dependences of the delay-time [95, 8] distributions $P(\Delta t)$, and the circadian cycles [9] in the time series $\{p(t)\}$ are obtained not just in the case of techno-social interactions. The presented ABM can work for a wide range of parameter values. In this Section we will use parameters extracted from the discussion-driven Digg data set in order to enable comparison of the simulation results with one obtained from the quantitative analysis of empirical data. As we will show, the simulations can be analyzed in a similar fashion as data obtained from the Blogs and Digg using methods presented in Chapter 2.
Some of the parameters, such as the relaxation rate of the arousal and the valence and the parameters d_1, d_2, c_1, c_2 of the maps in Eqs. (4.5-4.6) can not be extracted from this type of empirical data. The values shown above are chosen such that the fixed points of the maps do not fall to corner areas for typical values of the environmental fields occurring in our simulations, cf. Figs. 4.6 (left) and (right).

4.2.4 Simulated temporal patterns

The system is driven by adding $p(t)$ new users at each time step and letting them to boost the activity of the system, according to the model rules introduced in Section 4.2.2. We sample different quantities in analogy to those that we can define and compute from the empirical data, see Chapter 3. Compared with the empirical data, the advantage of the agent-based model is that we can also track of the fluctuations in the valence and the arousal of each agent ("user") at all time steps.

Typically, the arousal and the valence of an agent, who is linked through the posts to other agents, experiences stochastic inputs from the active environment, as

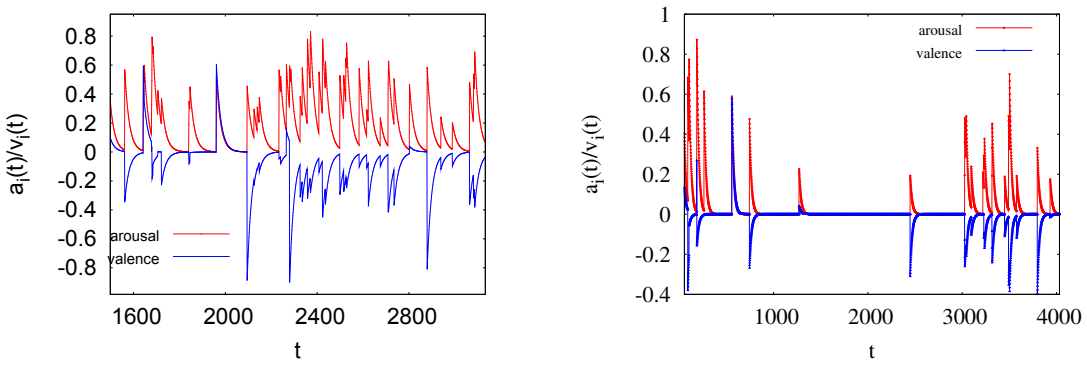

Figure 4.9: Two examples of the valence and the arousal are shown against time for two agents (users) located in different areas of the bipartite network, resulting in different activity patterns: a very active agent (left) and a sporadically active agent (right).

shown in Fig. 4.2.4. Between such events the emotion arousal and valence decay with the rate γ towards zero values. It should be stressed that at each agent (user-node) different patterns of the activity are expected. They depend not only on the current network structure surrounding the agent, but also on the fact that at given time the activity might be transferred to another part of the network, i.e., due to the aging of posts and the preferences of other agents towards particular types of posts. Two illustrative examples shown in Fig. 4.2.4 are from the same simulation run, but for two agents who are located at different areas of the network.
The dynamics of emotional states of users depend on several model parameters. For example, the relaxation rate, γ, determines how long the user will stay in emotionally excited state if he is not exposed to the environment, i.e., he is not present on the blog. Slow relaxing emotional state enables a user to get involved more often in a discussion type of the activity, and consequently having increased effects to the environment. Whereas, the users with fast relaxing states remain mostly inactive between the occasions where they receive inputs from the environment. The Fig. 4.10 shows fluctuations of valence and arousal for two users of different activity, with relaxation rate $\gamma = 0.3$, and values of all other parameters the same as for slow-relaxation case. Emotional states of both of these users relax faster than in the case when $\gamma = 0.05$ in Fig. 4.2.4.

The actions of individual agents contribute to the overall activity that can be monitored at each post and at the whole (evolving) network, as well as at the network parts, for instance the topological communities, that can be identified when the network is large enough. In the simulations we monitor the fluctuations of the number of active posts, $N_{ap}(t)$, the number of different agents that are active at these

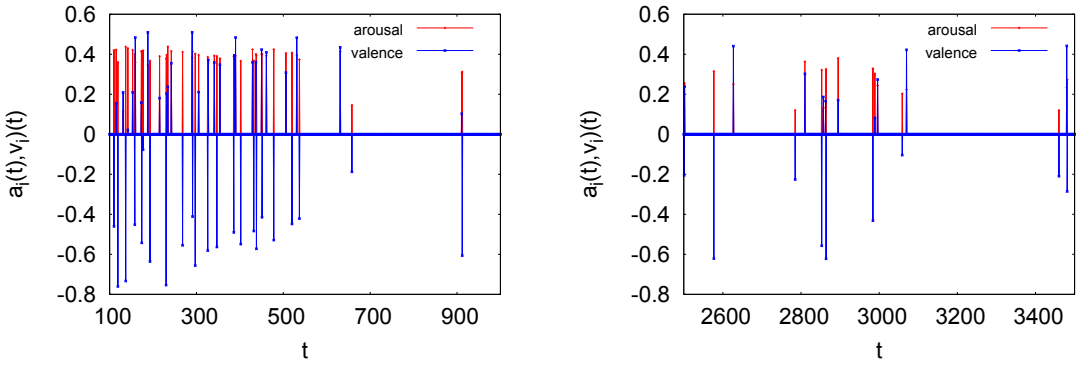

Figure 4.10: Fluctuations of emotional state of a very active (left) and sporadically active (right) agent with relaxation rate $\gamma = 0.3$

posts, $N_{au}(t)$, and the number of comments that these agents posted at each time step, $N_c(t)$. Furthermore, we distinguish between the comments that carry positive (negative) valence, $N_{\pm}(t)$, and the overall charge of these emotional comments, $Q(t)$. The temporal fluctuations of these quantities are shown in Fig. 4.11 (a) and (b), where only the initial part of the time-series are shown, corresponding to four weeks of real time. Notice, the circadian cycles of the driving signal are reflected to the time-series of the number of active agents and the number of their comments.

The power spectra $S(\nu)$ of these time series are shown in the upper panels in Figs. 4.11 (c) and (d). A characteristic peak corresponding to the daily cycles of the time-series is visible. In addition, long-range correlations with $S(\nu) \sim 1/\nu^{\phi}$ occur in most of these time series (except for the charge fluctuations!) for the range of frequencies, indicated by the slopes of the straight lines in both figures. The simulated time-series can be compared with the ones observed in the empirical data, see Chapter 3. The fractality of these time series, leading to the power spectrum of the type $1/\nu^{\phi}$, as well as the dominance of the negative charge suggest that our model captures the basic features of the blogging dynamics.
Specifically, in response to the same driving signal, which has the power spectrum with the exponent $\phi = 1.5$, the simulated blogging process builds the long-range correlations yielding the time-series with smaller exponents $\phi = 1.33$ and $\phi = 1$, in the number of active agents and the number of comments, respectively, and increased range of correlations, qualitatively similar to the popular Digg.
One could argue that correlation in time series obtained from the simulations are induced with the time series $p(t)$. In order to show that long-range temporal correlations are inherent to the stochastic process of our model, we simulate system dynamics for $p(t) = 6$ with the same values of all other parameters as in the previous case. The simulated time-series of number of comments and charge are shown in Fig. 4.12 (left). Apart from higher average activity and the absence of daily cycles,

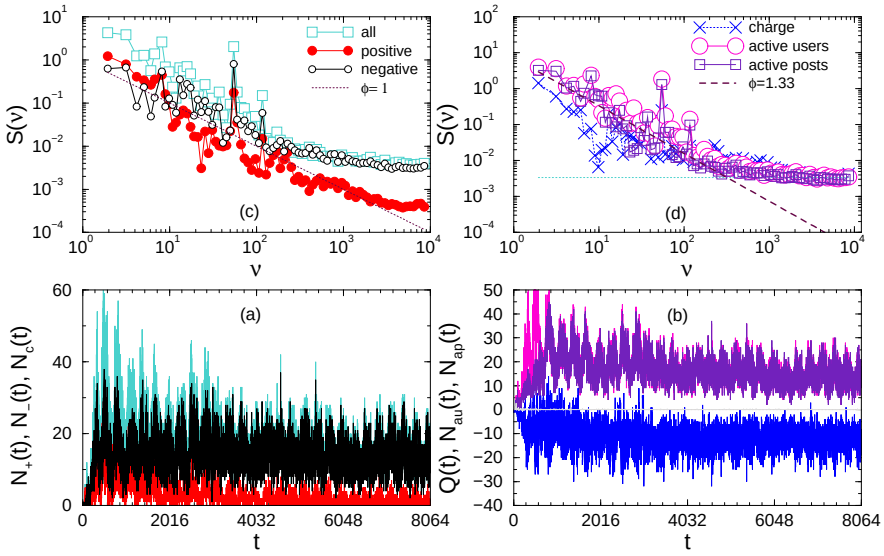

Figure 4.11: The number of all comments per time step (cyan) and the number of comments with positive (red) and negative (black) valence per time step (a), and the double-logarithmic plot of their power spectra (c). The number of active posts (indigo) and the number of active agents (magenta) per time step, and the charge of all emotional comments (blue), are shown in panel (b), and their corresponding power spectra, panel (d). Straight lines indicate slopes ϕ={1, 1.33, 0}. For clear display, the power-spectra are logarithmically binned.

the fractality of the time-series is preserved. The power-spectrum with the exponent $\phi \approx 1$ is observed for the number of comments, although in a smaller range. The negative charge sets-in after certain time period and fluctuates in a stationary manner. This shows that long-range correlation can be enhanced but not imported by the profile of the driving noise.

4.2.5 Simulation of emergent bipartite networks and communities of the emotional agents

Further we investigate simulations at the level of the emergent network topology. The networks that emerge through the activity of our emotional agents on posts can be studied in full analogy with the bipartite networks mapped from the empirical data of the same structure (see Chapter 3). In our simulations the network evolves due to the addition of nodes of both partitions, as well as the evolution of links. The lists of user - posts connections are updated at every time step. With the prevailing negative comments, as demonstrated above in the time-series, cf. Fig. 4.11,

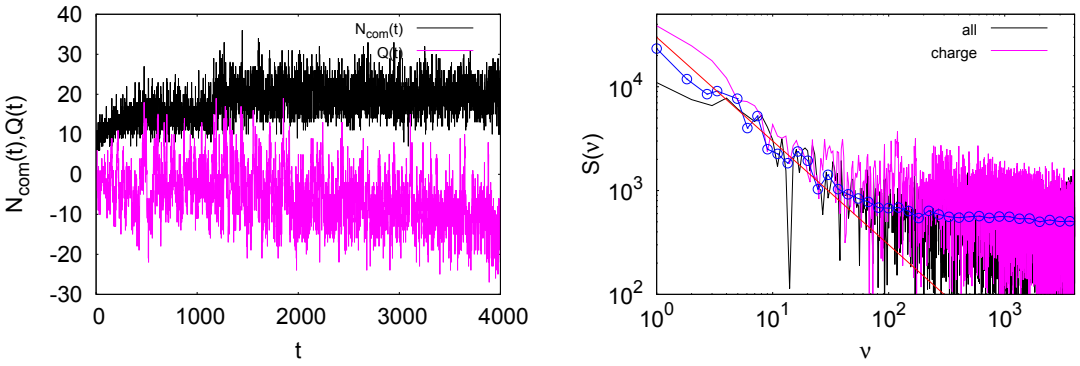

Figure 4.12: Time-series of the number of comments and charge (left) and their power-spectra (right) obtained when the system of the emotional agents is driven by adding a constant number of users $p = 6$ per time step.

arrivals of the negative comments at posts generate an environment that, in view of the linking rules (large negative charge preference), strongly affects the network evolution. In this way some posts with a large number of negative comments and thus large topological strength may appear.
A part of the emergent network obtained in our simulations is shown in Fig. 4.13, where three such hubs-popular posts are visible together with the users linked to them. Some topology measures of the emergent bipartite network-the degree distributions and the assortativity measures, are shown in Fig. 4.14 (a) and (b). As expected, the degree distributions for each of the partitions-agent(user)-nodes and post-nodes appear to be different. Specifically, the broad distributions are dominated by different type of cut-offs. They can be approximated by the following mathematical expressions, motivated by fitting the corresponding empirical data:

$$P(q_u) = C_u q_u^{-\tau} e^{-q_u/X_{u0}}; \qquad P(q_p) = C_p \left[1 - (1-q)\left(\frac{q_p}{X_{p0}}\right)\right]^{1/1-q}; \qquad (4.10)$$

for the agent(user)-node and the post-node distributions, respectively. The agent-degree distribution $P(q_U)$ exhibits a short power-law region and a large exponential cut-off, whereas the distribution related to the post-nodes $P(q_P)$ has a dominant cut-off at smaller degree followed by a power-law tail, compatible with the q-exponential form with $q > 1$. The fitted values of the parameters are shown in the Figure legends in Fig. 4.14(left). In principle, they depend on the simulation parameters of the agent-based model. However, the expressions appear to be stable with respect to the simulation time (size of the network). The results are shown for two simulation runs with 16384 and 25000 time steps, resulting in the networks with $N_P = 13504 + N_U = 64852$, and $N_P = 22757 + N_U = 107933$ nodes, respectively.

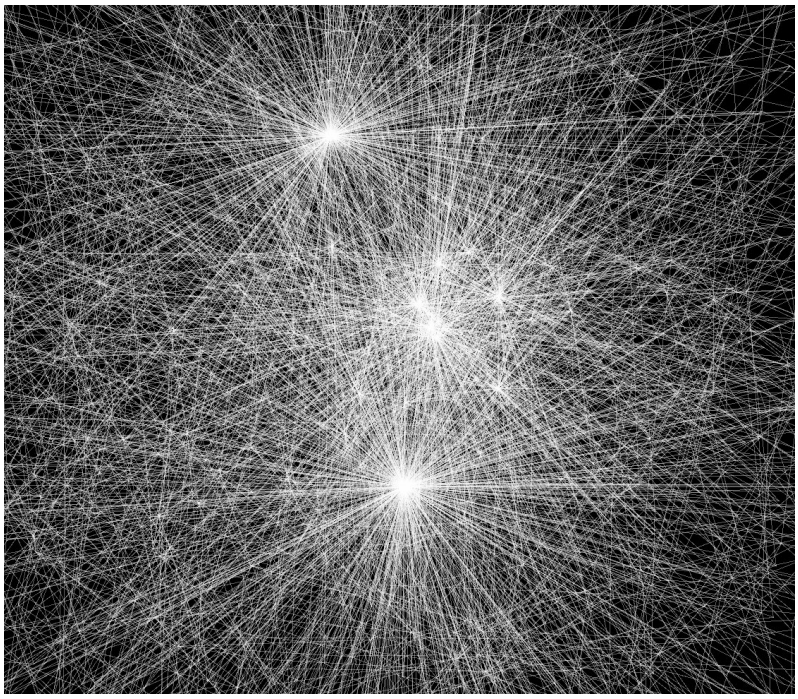

Figure 4.13: Shown is a part of the network structure in the vicinity of three popular posts.

As it was shown in Chapter 3, the same mathematical expression Eq. 4.10 apply (with different exponent) to corresponding distribution derived from Digg data set. These topological measures demonstrate that bipartite network emerging in the emotional blogging of our model shares qualitative similarity with the one from the popular posts in real data. To certain extent, similar conclusions apply in the case of the mixing patterns, shown in Figs.4.14 (right) and 3.5 (right). The mixing patterns are calculated according to Eq. 2.2, i.e., we calculated normal degree-degree correlations. The posts making the network neighborhood of an agent-node (user-node in the empirical data) exhibit no assortativity measure, which is indicated by the line of zero slope before a cut-off at large user-node degree. In the case of post-nodes, the empirical data indicate slight disassortativity (decrease) just before the cut-off, Fig. 3.5 (right), a feature that seems not be properly captured by our model in the present parameter range.
The model parameters can have different effects on systems features such topological properties of bipartite network and its projections or activity patterns. Hear we show how change in parameter values for $p(t)$ and T_0 ($\mu(T_0)$) influence network topology.
Figure 4.15 indicates that the properties of driving force do not influence the topological features of obtained bipartite network. The obtained network consists of

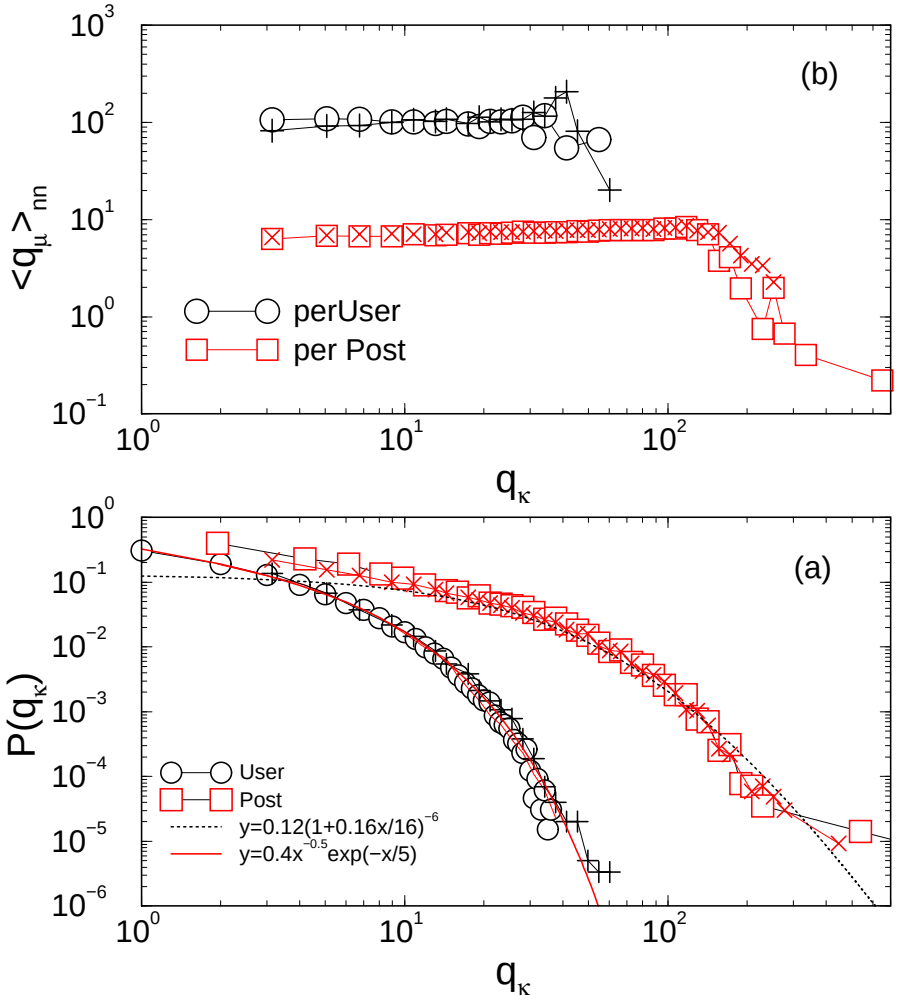

Figure 4.14: (a) The degree distributions of the user-partition (◯,+) and the post-partition (□,×). Fitting lines explained in the Legend. (b) Assortativity measures: The average degree of the posts linked to the user node of a given degree versus user degree (◯,+), and the average user degree linked to the post of a given degree, plotted against post degree (□,×). Empty symbols are for the simulation time 16384 steps, while the crosses indicate the respective results from runs with 25000 time steps.

$N_U = 98304$ connected to $N_P = 18777$ posts. Users in this simulated system posted $N_C = 314344$ emotional comments during a period of 16384 time steps. The degree distributions in user and post partition are of the same type as Eq. 4.14. In order to investigate the mixing patterns for bipartite weighted network obtained for constant $p(t)$, we calculate weighted average nearest neighbor degree according to Eq.

2.6. As in the case of the network for fluctuating $p(t)$, the degrees of nodes in both partitions are not correlated 4.15 (right).

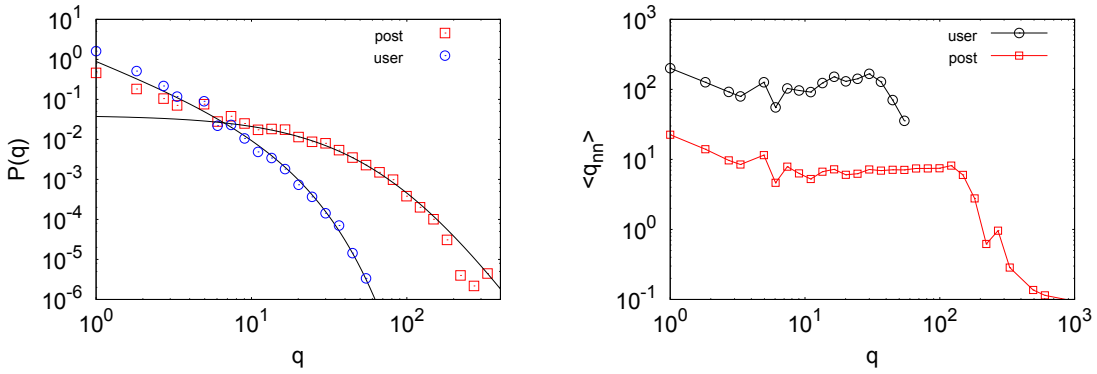

Figure 4.15: Degree distribution (left) and mixing pattern (right) for user and post partitions of bipartite network obtained from the simulations for $p(t) = 6$.

The number of comments on a certain post strongly depends on its exposure time. During the period of its exposure, determined with time window T_0, the probability that post will be commented is proportional to the number of its comments. After the post becomes old, the probability that it will gain a comment depends only on the value of its negative charge. Dynamics on exposed and old posts differs drastically. It follows from this, that dynamics of the systems, so as network topology, depends on the value of the parameter T_0. In one of the limiting cases, when $T_0 = n_t$, all posts are constantly exposed and only preferential dynamics is present. The degree distribution of posts will depend only on $P(t_{tp})$ and will exhibit power-law behavior. Hear we show the results for $T_0 = 2016$ (period of one week) for which the dissemination parameter has the value $\mu = 0.012$ (see Fig. 4.8). The bipartite network for the period of $n_t = 16384$ time steps is of size $N_U + N_P = 64852 + 13841$. The change in T_0 (μ) influence activity in user partition which is reflected in smaller maximal degree, but the type of distribution, shown in inset of Fig. 4.16 (left) remains the same. On the other hand, the out degree distribution of posts exhibit different behavior for larger values of degree Fig. 4.16 (left). The mixing patterns are also influenced by the change of exposure time of the post. The weighted average nearest neighbor degree of agent decays with its degree, indicating disassortative mixing in the user partition (inset Fig. 4.16 (right)). The disassortative mixing was also observed in post partition for small and large values of post degree, while for degrees between 10 and 100 the degree correlations do not appear.

The existence of communities in bipartite networks and their projections are is an indicator of collective behavior. We are interested in communities of the emotional agents, that may potentially occur on the respective monopartite projection of the weighted bipartite network.

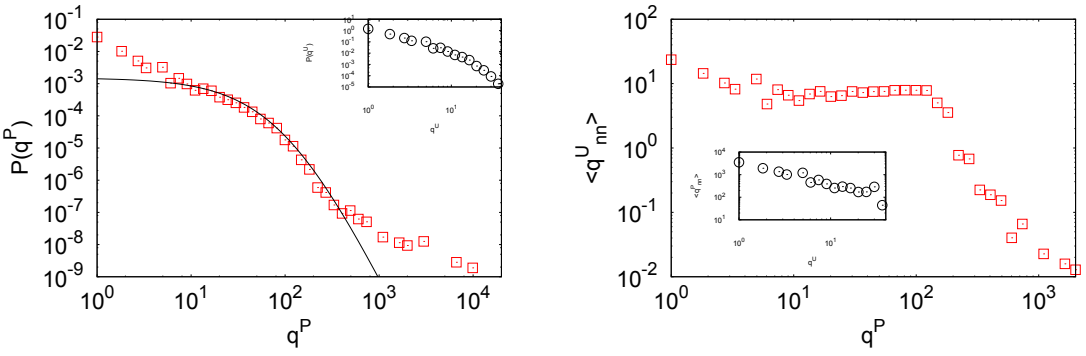

Figure 4.16: (left) Degree distribution and (right) mixing patterns in post partition for weighted bipartite network, obtained from simulations for T_0. In insets are shown degree distribution and correlations in user partition.

The monopartite user- or post-projections of these networks obtained from empirical data, appear to be highly clustered and weighted networks, Chapter 3. The networks emerging through the actions of the emotional agents in our simulations have similar features, for which reasons we find that eigenvalue spectral analysis method is suitable for identification of their community structure (Section 2.1.2). We perform the spectral analysis of the normalized Laplacian operator which is related to the weighted user-projection network, whose matrix elements C^W_{ij} represent the common number of posts per pair of users, *including the multiplicity of user–post connections*, which is indicated by the superscript. It is constructed from the symmetric matrix of *commons* as

$$L_{ij} = \delta_{ij} - \frac{C^W_{ij}}{\sqrt{(l_i l_j)}} \ , \tag{4.11}$$

where l_i is again the *strength* of node i. The network grown through the emotional actions of the agents in our model has specific properties which may be reflected in the community structure. Namely, the bipartite network is already weighted, which shifts the distribution of weights in the monopartite projection away from pure topology of commons [17]. In addition, the network evolves in such way that the center of the activity is shifting to ever new groups of (exposed) posts.
Here we analyze the structure of the network after 4032 time steps (two weeks) of the evolution. The network projected onto user (agents) partition contains $N_U = 4572$ users, only users with the degree larger than 5 are considered as relevant for the community formation. The eigenvalue spectrum of the Laplacian operator Eq. 4.11 with the C^W_{ij} matrix related with this user-projected weighted network, is then computed. The results for the eigenvalues shown in the ranking order are given in Fig. 4.17 (left). The scatter-plot of three eigenvectors belonging to the three lowest

eigenvalues is shown in Fig. 4.17 (right). The spectrum, as well as the scatter-plot in Fig. 4.17 (right), indicate that five agent-communities can be differentiated. These are denoted by G_k, with $k = 1, 2, \cdots 5$ corresponding to top-to-bottom branches in Fig. 4.17 (right). In the following we first identify the nodes representing the agents in each of these communities. Then we analyze how the communities actually evolved on that network and discuss the fluctuations of the emotional states of each agent in the communities through the evolution time.

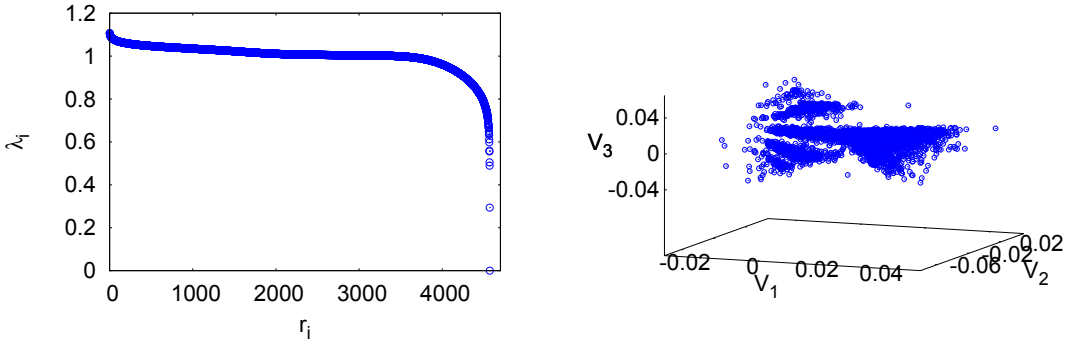

Figure 4.17: For the agent-projected network (left) the eigenvalue spectrum, and (right) the scatter-plot of three eigenvectors belonging to the lowest nonzero eigenvalues, indicate five communities G_k, $k = 1, 2 \cdots 5$, related to five branches in the scatter-plot, from top to bottom.

Quantitative analysis of the empirical data shown in Chapter 3 suggest that the communities of posts often appear in relation to their subjects, age, and sometimes authorship. The features is prominent in the case of Blogs of "normal" popularity. Whereas, in the communities on popular posts the subjects are often mixed, leaving potentially different driving force for user's intensive activity on that posts. In our model the post have no defined subjects, and the agents are attracted by the emotional content of the post/comment. As we will show in the following Section the emotions play an important role in users collective behavior.

The spectrum and scatter plot for projected network of users obtained for constant p, show that the type of driving force influence the size and the number of communities but not their dynamics. We reduced the size of users partition to $N_U = 4418$ by selecting only agents that made more than 10 comments. The monopartite network of these users is created by calculating matrix of weighted commons C^W, where element C^W_{ij} represents the number of posts on which both users left a comment. The eigenvalue spectral analysis of Laplacian 4.11 for this network, reveals the existence of three communities in a network of very active users 4.18. It will be shown in the following Section, that the size and dynamics of

these communities is closely associated with emotions expressed by their members.

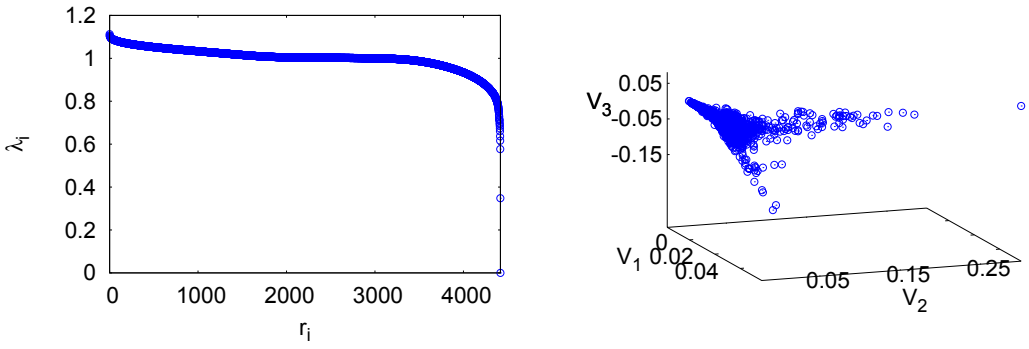

Figure 4.18: Spectrum (left) and scatter plot (right) of Laplacian matrix for agent-projected network for $p(t) = 6$. The branched structure of eigenvectors for the three smallest non-zero eigenvalues indicates existence of three user communities, G_k $(k \in 1, 2, 3)$.

4.2.6 Temporal patterns of emotional communities

Having identified the agents in each of the communities, we can track of their group activity and the emotion fluctuations over time from our simulation data for both networks.

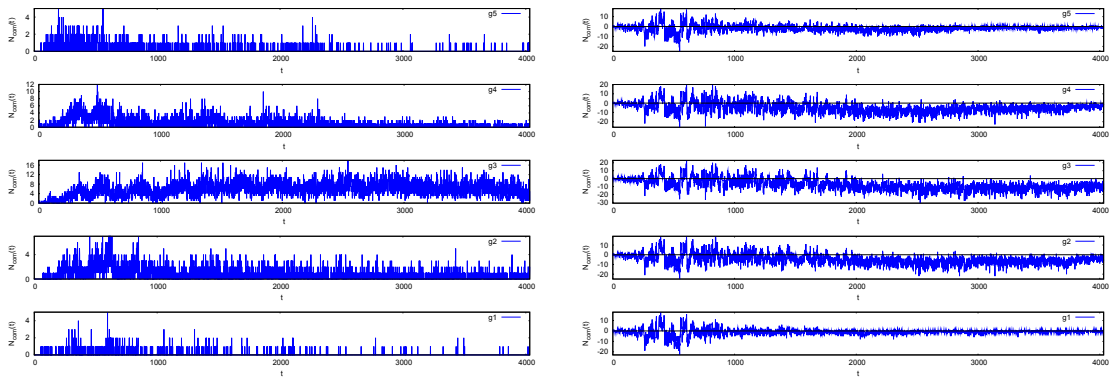

Figure 4.19: Time series of the number of comments by the agents belonging to a given community G_k (left), and the charge of these comments (right), computed for each of five communities identified in Fig. 4.17.

The time-series of the number of comments of all agents in a given community G_k are shown in Figs. 4.19 and 4.20 (left) and the emotional charge of the valence

of these comments in the Figs. 4.19 and 4.20 (right). Note that a fraction of comments with the valence values close to zero in the range between $(-0.01, +0.01)$ are considered as neutral, and do not contribute to the charge.

The profile of the time series shown in Fig. 4.19 (left) indicates that all communities started to grow at early stages of the network evolution. However, two of them, G_1 and G_5, ceased to grow and reduced the activity relatively quickly after their appearance. Looking at the fluctuations of charge of the emotional comments in these two communities, we find that it is well balanced, fluctuating around zero at early times, and eventually leveling up to zero. Whereas, in the other two medium-size communities, G_2, G_4, the activity is slowly decreasing, while the largest central community, G_3, shows constantly large activity. Comparing the activity (number of comments) with the fluctuations in the charge of the emotional comments, we can see that in these three communities the excess negative charge settles after some time, breaking the initial balance in the charge fluctuations.

The communities found in network for constant $p(t)$ have slightly different temporal behavior Fig. 4.20 (left). The activity in all three groups is present from the beginning, but their activity peaks occur at different time moments and intervals. Specifically, group G_1, which is relatively small, has the highest activity in first 2000 steps (approximately period of one week), after which the activity of its members becomes sporadic. Other two groups, G_2 and G_3, experience their maximal activity after a number of simulation steps which corresponds to period 2 weeks in real time. Although, the number of comments posted per time drops, users belonging to these two communities are constantly active. The charge per time bin in communities G_2 and G_3 becomes negative with the increase of the number of comments, Fig. 4.20 (right), while the charge for the group G_1 is balanced.

In this way our model reveals the correlations between the prolonged activity and

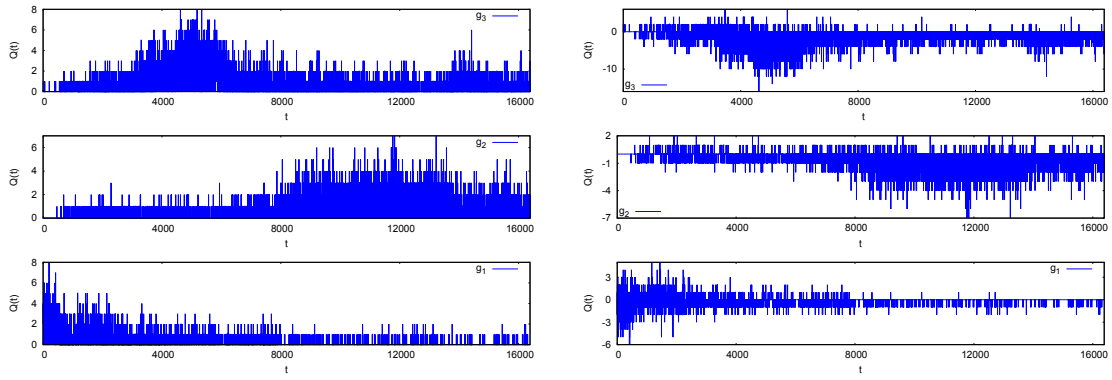

Figure 4.20: Time series of number of comments (left) and charge (right) for three communities identified in Fig. 4.18

the size of a community (i.e., number of different agents), on one side, with the

occurrence of the negative charge of the related comments, on the other, a feature also observed in the empirical data on Blogs and Digg.

4.2.7 Circumplex map

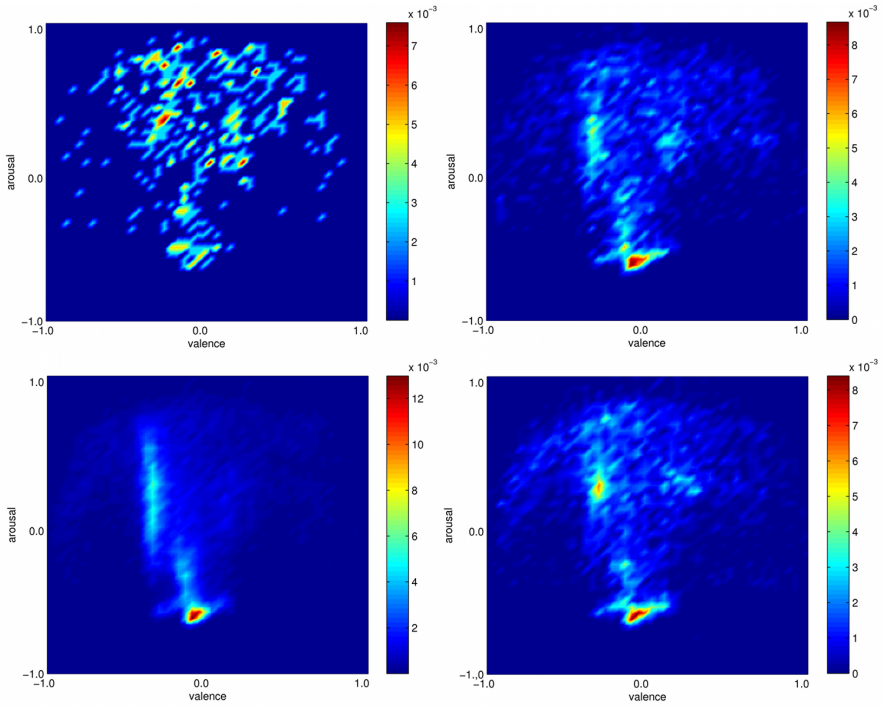

Figure 4.21: Circumplex map of the emotional states of the agents belonging to four communities identified on the emergent network: G_1–top left, G_2–top right, G_3–bottom left, G_4–bottom right. Color map indicates occupancy of a given state, normalized relative to the number of comments in each community.

In view of the preference towards the posts with negative charge, a comment regarding the breaking of the charge balance and its consequences to the network topology is in order. Note that, according to the model rules, the probability of a post to receive first negative/positive comment depends on the valence of the agent who is active on that post and its similarity with the average valence of the currently active posts. Contrary to the naive preference towards node's degree, which is known to lead to a scale-free degree distribution in growing monopartite networks (for theoretical derivation and the conditions when power-law distributions occur, see Ref. [92]), in our model agents preference is driven by another quality of a post node, its emotional content. Hence, in this process no scale-free distributions of the posts degree is expected, as also shown above. More importantly, the negative

charge appears to have limited fluctuations. The time-series of the (negative) charge and of the number of comments remains stationary over large periods of time, before the activity ceases, as shown in Figs. 4.19 and 4.20 for the communities, and in Fig. 4.11 (b) for the entire network.

Another interesting feature of these communities can be observed by visualizing the patterns of their activity in the *phase space of the emotion variables*. In order to match the emotion measures accepted in the psychology literature, i.e., according to the 2-dimensional Russell's model [112, 120], we use the circumplex map explained in Section 3.1.3. The values of the arousal and the valence variables are thus mapped onto a surface enclosed by a circle, where, according to Refs. [120, 112], the emotions commonly known, as for instance, "afraid", "astonished", "bored", "depressed", etc., can be represented by different points (or segments) of the surface.

Computing the transformed values of the arousal and the valence for each agent in a given community at all time steps when an action of that agent is recorded in our simulations, we obtain the color-plots shown in Figs. 4.21. Specifically, the color map indicates how often a particular state on the circumplex was occupied in four of the above communities, normalized with the all actions in that community. As the Figure 4.21 shows, the communities leave different patterns in the space of emotions. For instance, the community G_1, that have balanced charge fluctuations, appears to cover a larger variety of the emotional states, leading to the pattern on the top left figure. Whereas, when a large community is formed, it may induce large negative fields which keep the agents in the negative valence area of the circumplex map. The situation corresponding to the community G_3 is shown in bottom left plot in Fig. 4.21. Majority of the comments is this case are centered in the area of the arousal and the valence where the negative emotional states known as "worried", "apathetic", and "suspicious", "impatient", "annoyed" etc, are found on the circumplex map (see Refs. [120, 114] for coordinates of some other well known emotional states covered by these patterns). Plots on the right-hand side of Fig. 4.21 correspond to the communities G_2 and G_4, in which charge fluctuations are moderately negative, as discussed above.

From the Figures 4.21 we can also observe that the well defined lower bound for the agent's arousal emerges in a self-organized manner inside each community, although no sharp threshold exists in the model rules. Moreover, the arousal drives the valence when the agents are active. This is clearly displayed in the case of communities with a balanced charge, as our community G_1, Fig. 4.21 top left. Similar pattern of the arousal–valence was found in the laboratory experiments [123], where the values of the arousal and valence are inferred from skin conduction, heart beat and facial expression measurements on users reading a selection of posted texts.

The characteristic patterns in Figs. 4.21 emanating from the emotional blogging of our agents suggest the processes with anomalous diffusion, in which certain parts of the phase space are more often visited than the others. They reflect the self-

organized dynamics of the agent's emotion variables and the network topology. Formally, these patterns are between two extreme situations: synchronous behavior, focusing at a lower-dimensional areas, and random diffusion, spreading evenly over the entire space.
The most visited areas of the phase space are in the vicinity of the attractors of the nonlinear maps. The positions of these attractors for each agent map move, depending on the values of parameters c_2 for valence and d_2 for arousal and on actual values of the fields acting on it. The fields themselves fluctuate over time for each map, being tunned by the the agent's emotion variables and the local topology of the network (community) where the agent is situated.
The agents in our model have random values of valence and arousal when they are

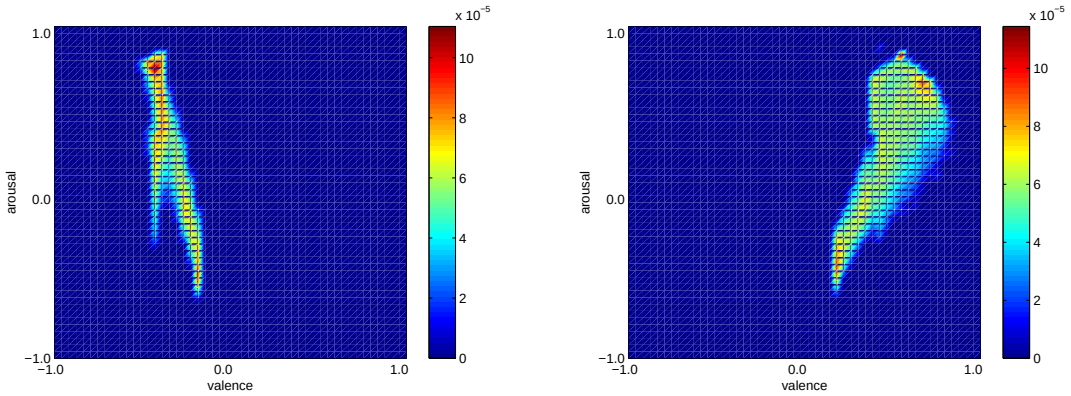

Figure 4.22: Circumplex map of the emotional states of the agents obtained from simulations if all agents are angry (left) or astonished (right) when they are added to the system.

introduced into the system. It is believed that this corresponds to a real situation, i.e., different people in different ways contribute to the overall emotional state of the virtual society. The dynamics in our system is driven solely by adding new users, meaning that the information from the *outside* gets into the system only through the initial values of emotional state of the agent. We test the features of the system in relation to the initial values of agent's arousal and valence (the values of all other parameters are given in Section 4.2.3): (i) user is *angry*, (ii) user is *astonished*, or (iii) user is either *angry* or *astonished* with equal probability, when it joins the system. The patterns in the space of emotions for case (i) and (ii) are given in Fig. 4.22. The anger is very aroused emotional state with moderate negative valence (the valence is around value -0.45). Starting from the beginning agents write only comments with negative and neutral emotional content, meaning that only negative field is present. Since there is no positive field, the valence map Eq. 4.6 has only negative atractors resulting in a pattern shown in Fig. 4.22 (left). On the other hand, when

all users start as astonished, emotion with very high arousal and valence around 0.45, comments can only be positive or neutral. Lack of negative comments and fields leads to pattern Fig. 4.22 (right), where most of the users are *astonished*, *ambitious*, *feeling superior* and *convinced*.

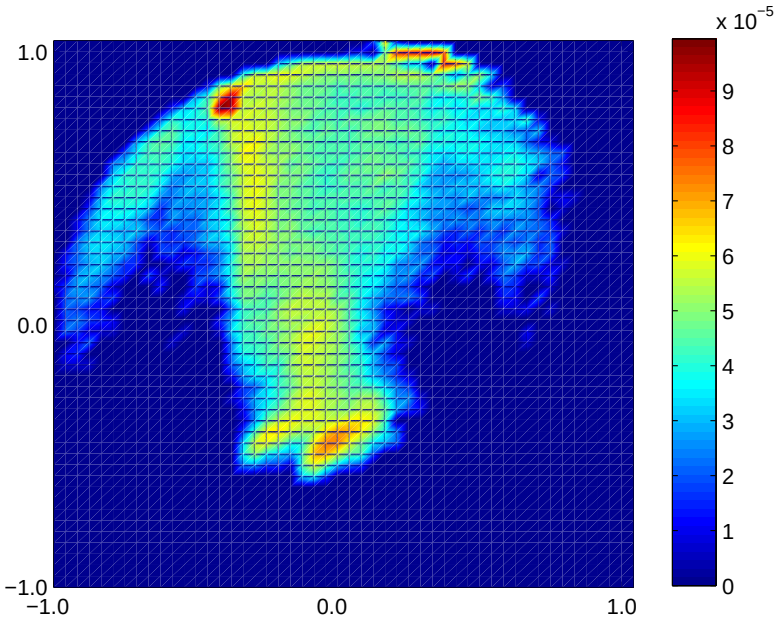

Figure 4.23: Circumplex map of the emotional states of the agents when their initial value of emotional state can be characterized either

In the third case, when added user is either astonished or angry, the pattern resembles the one obtained for users with random initial values of arousal and valence (see Fig. 4.23), indicating that the dynamics of emotions in the system depends strongly on the parameter c_2.

Chapter 5

Summary and Conclusions

The emergence of the web and rich social computing applications enable physics of complex systems a unique opportunity to study complex human behavior powered by emotions. By studying users behavior on the Blogs and similar sites we are able to observe and model influence of emotions on collective behavior on the Web. This knowledge should help us to better understand not just human on-line social behavior but also off-line which is usually hard to study due to a luck of data. For this reasons it is important to understand how on-line societies work, what is their structure, and to understand the global consequences of our micro-level decisions. Especially we are interested in understanding how emotions influence our on-line actions, our interactions with other users, and overall global behavior that emerges from it. In this thesis we presented methodology for quantitative study of on-line interactions, combined of empirical work, collection and analysis of the data, and theoretical modeling of techno-social systems.

The research focus of this thesis is to analyze and model the structure, evolution and dynamics of Blogs and similar websites. We showed how the methods of *complex network theory* and *statistical physics of complex systems* can be used for studying system with computer-mediated interactions. Using frameworks of cellular automaton models and agent-based modeling we investigate the role of emotions in the observed emergent phenomena on Blogs. Our main contribution is development of methodology for the quantitative analysis of empirical data and simulations of techno-social interactions. We showed how the data from Blogs and similar websites can be mapped onto bipartite networks. How the analysis of the structure of these networks and their projections can be used for diagnostication of collective behavior. Trough the statistical analysis of Blogs and Digg data sets we investigated the influence of the emotions on Blog dynamics, and showed. Using data driven models we were able to investigate this role independently of other features characteristic for the humans.

5.1 Network properties and emotions

Complex systems can be studied using machinery of complex networks and statistical physics (see Chapter 2). The constituents of the system can be represented with nodes while interactions are represented by links connecting them. The nature of the nodes and links, which determines the network type, depends on the system features and on the scale at which we want to study its behavior. The structure of the network strongly depends on system dynamics and function. Different topological properties of networks can be used for analysis and understanding of complex behavior of the system. The degree distribution and mixing patterns describe the systems properties on global scale while community structure can provide information about clustering processes in the networks. Although, studied Blogs and Digg are social systems, the obtained networks exhibit some properties characteristic for technological systems. Specifically, they exhibit disassortative mixing patterns, with a tendency to have connections between vertices with dissimilar degrees. This kind of behavior, found also in the network of the Web [35], indicates that interactions on the Web can not be fully regarded as purely social but that the impact of technological networks should be also considered.

The analysis of Blog and Digg networks (see Chapter 2) shows high heterogeneity within both partitions, users and posts. In fact, based on degree distribution posts can be divided into two groups: *normal* and *popular* post. The dynamics on this two distinguished groups differs and depends on different user properties [12, 16, 17, 18]. Analysis of projected user networks related to these groups showed that on Blogs the users appear to be normally clustered around preferred subjects. However, mixing between subjects increases when a post becomes popular, suggesting another active principle, which can be found by a direct inspection of the posted material. User communities occurring in relation to popular Blogs can be readily identified and their structure and evolution analyzed 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. 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. Specifically, we showed that presence of criticism in comments connect community more tightly by instigating its activity. This is opposite to situations present in real world and standard on-line social networks where positive emotions are more preferable [124].

Further, the role of emotions in the formation of communities is investigated using agent-based model of emotional blogging (see Section 4.2). The main property of each agent is its emotional state, measured with two dynamical variables of arousal and valence. The activity of agent, so as polarity its actions, depends on these two variables. Beside this variables, the dynamics of model also depends on frequency of appearance of new agents in the network, which is characterized with circadian

cycles, and time-delay parameter, both inferred from the empirical data. The agents spread emotions via indirect contacts through their actions on the posts. Together with agents and their connections, these posts are forming bipartite network which evolves in time due to addition of agents and their actions. By changing different control parameters in the model one can investigate system dynamics and its dependence on: (a) properties of each individual agent; (b) on linking rules (c) driving force of the system and the balance between the influences of local and global events. In Section 4.2 we demonstrated appearance of agent communities which emerge only as a consequence of the emotional commenting of the posts. These communities are identified as topological subgraphs in weighted-projection of bipartite weighted network on a user partition. The analysis of evolution of these communities reveal that agents are mostly active in their community. Their growth is self-amplified and prolonged with the excess negative charge of the related comments.

5.2 Self-organized criticality of emotional behavior

In this thesis we showed how the methods and theoretical concepts of statistical physics of complex systems can be used for quantitative analysis of empirical data and simulations. The analysis of the data from Blogs and Digg, indicates that the temporal behavior of users on Web, described trough distribution of time delays, is independable of the site. Moreover, it was found that the characteristic behavior doesn't depend on the type of the communication environment [95, 12, 16] but it is rather a consequence of some general features of humans. For this reasons, the time delay of agent's action, must be derived from the data and included as a external parameter in the model of human emotional dynamics on the Web. As we showed in Section 4.1, the dynamics of the system depends on the type of the distribution of time delays (for details see Section 4.1 and Ref. [18]).
The activity of the users can be described as a time series of number of comments. The analysis of different time series reveals the existence of long-range temporal correlations, with superimposed circadian, daily and weekly cycles. The long-range correlations are manifested in power-spectrum of time series. In fact, power-spectra of time series of number of all, negative, positive and emotional comments, so as number of active posts and users exhibit $\nu^{-\phi}$ behavior for small frequencies (long time intervals) with the exponent $\phi \in [-1, -1.5]$. This shows that user's activity is correlated for large time scales. The analysis of these time series obtained from simulations of ABM show that this correlations can be manifested in pure emotional blogging, demonstrating the important role of emotions in collective behavior of Web users.

Further, the collective behavior of users in both, empirical data and simulation, can be quantitatively measured in therms of connected avalanches, Chapter 4 and Ref. [18, 125]. We determine several quantities to characterize SOC structure of these avalanches. Distribution of avalanches sizes both on the entire bipartite network and for the truncated avalanches on single-post networks so as distribution of time intervals between the them and their duration, exhibit power-law behavior characteristic for self-organized critical systems.
Properties of this emergent collective state can be studied trough both of proposed models. Network automaton model proposed in Section 4.1 is suitable for studying the dependence of SOC emotional state on two crucial parameters: time-delayed actions and spreading of emotions. The analysis of simulations show that the observed critical states appear to be quite robust when the parameters of user behavior are varied within the model. However, they are prone to over reaction with supercritical emotional avalanches triggered by a small fraction of very active users, who disseminate activity (and emotions) over different posts. Swinging through a critical point at $\lambda = \lambda_c$, the collective states with emotional avalanches thus can be identified as having supercritical, critical or subcritical features. Note that in the empirical data the actual value of the parameter λ is not fixed but varies over time and communities considered. The simulations from ABM support the findings of network automaton model indicating the spreading of emotions as an important factor in emergence of critical collective behavior.

5.3 Future work

In this thesis we proposed a methodology for studying social interactions on the Blogs and similar Web sites. On this sites users do not communicate directly but trough posts and comments on them. We showed that emotions expressed in this comments have important role in emergent collective behavior. Beside Blogs, humans also interact trough different types of social networks, such as Facebook, MySpace, Google+, chat rooms and forums. Although, the interactions on these sites are trough messages, they differ a lot compared to one on Blogs. One of the most important features is lack of posts and the communication is direct. Users often address other users in their messages making this communication direct. For the analysis of this data, one can use the same methodology proposed in this thesis with some modifications. For example, the mapping of empirical data from these sites differs from one proposed for the Blogs. Due to slightly changed communication mechanisms in these social sites compared to one on Blogs, we expect different behavior. For instance, some analysis of MySpace data reveal that most of the communication is positive and for this reasons we assume that other principals govern dynamics in

this system. To study the role of emotions one needs to develop separate models that capture the dynamical characteristics of each of mentioned systems. The models of emotional agents, proposed in Ref. [118] and this thesis, represent the basis for these future models.
The scientific blogging has become an activity which involves not only scientists but also other people, who write on regular basis and are interested in science. The leading scientific journals, *Science* and *Nature*, have developed their own scientific Blog pages. These blogs are written by editors, journalists and members of their networks, and it is expected that posts on these Blogs meet high standards in terms of content and quality o writing. On the other hand on the various public services, such as *Blogger* and *Wordpress*, a huge amounts of content,some of which are related to science, are created every day. These posts are not subject of any evaluation process of its content, which is why the validity of the facts presented in them is questionable. As it was stated in Ref. [126] "we can no longer control who is writing about science" and that "perhaps we should ask whose science writing should we be reading". This raises the questions about the proper methodology for classification of scientific blogs based on the their quality. The methodology for quantitative analysis of empirical data non-scientific Blogs, presented in this thesis, can potentially be used for classification of scientific Blogs Ref. [126]. In this regard, future work in this area would involve crawling the data from the scientific blogs and analysis of their dynamics using the proposed methods, with special emphasis on the role of emotion in scientific blogging.

Appendix A

Eigenvalue spectral analysis method

To obtain eigenvalues and eigenvectors of normalized Laplacian matrix for weighted networks we use routines from Numerical recipes [127]. Specifically, we user routines *tred*2.*c* and *tqli.c*. The routine *tred*2.*c* is used for reduction of a symmetric matrix to tridiagonal form according to Householder reduction (for detail see Ref. [127]), while routin *tqli.c* is used for finding the eigenvalues and eigenvectors of tridiagonal matrices. The eigenvalue spectral analysis method consist of two codes: *tdm.cc* which gives tridiagonal form of normalized Laplacian matrix of weighted network, and *evl.cc* which is used to obtain the spectrum and eigenvectors of Laplacian matrix. The input file for the *tdm.cc* code is matrix of weights, W_{ij}, given in the form of a link list, i.e., every line contain the node numbers, v_1 and v_2, and the weight of the link between them lw. The matrix of weights is then transformed into normalized Laplacian matrix according to Eq. 3.14. The tridiagonal form of Laplacian matrix is then found using routine *tred*2.*c*. The output of the code *tdm.cc* is the arrays of diagonal (*diag*) and subdiagonal (*subdiag*) elements of tridiagonal matrix, and tridiagonal form of matrix (*tdmatrix.data*).

```
/*tdm.cc*/

#include <math.h>
#include <stdio.h>
#include <stdlib.h>
#include <iostream>
#include <fstream>
#include <sstream>

using namespace std;
```

```
extern double tred2 (double **a, int n, double d[], double e[]);

double **Ain;
double *d;
double *e;
double *q;
int lw,n,i,j,v1,v2,nmin;
float pom;

main (int argc, char **argv)
{

n=atol(argv[1]);

FILE *f,*h,*h1;
f=fopen("tdmatrix.txt","a");
h=fopen("diag","a");
h1=fopen("subdiag","a");

/*memory allocation*/
Ain=new double*[n+1];
d=new double[n+1];
e=new double[n+1];
q=new double[n+1];
for(i=0;i<=n;i++) Ain[i]=new double[n+1];

/*matrix input*/
for (i=1;i<=n;i++)
  {
   for(j=1;j<=n;j++)
    {
    Ain[i][j]=0;
    }
  }
while(!cin.eof())
  {
  cin >> v1 >> v2 >> lw;
  Ain[v1][v2]=-1.0*lw;
  }
for(i=1;i<=n;i++)
  {
```

```
  Ain[i][i]=0;
  }
for(i=1;i<=n;i++)
  {
  sum=0.0;
  for(j=1;j<=n;j++) sum=sum+(-1.0)*Ain[i][j];
  q[i]=sum;
  if(q[i]==0) k++;
  }
for(i=1;i<=n;i++)
  {
  for(j=1;j<=n;j++) if(!(q[i]==0) && !(q[j]==0))
Ain[i][j]=Ain[i][j]/(sqrt(fabs(q[i])*fabs(q[j])));
  Ain[i][i]=1;
  }

tred2(Ain,n,d,e);

for(i=1;i<=n;i++)
  {
  for(j=1;j<=n;j++)
    {
    fprintf(f,"%f ",Ain[i][j]);
    }
  fprintf(h,"%f\n",d[i]);
  fprintf(h1,"%f\n",e[i]);
  fprintf(f,"0 \n");
  }

for(i=1;i<n+1;i++) delete [] Ain[i];
delete [] Ain;
delete [] d;
delete [] e;
return EXIT_SUCCESS;
}
```

The files *diag*, *subdiag* and *tdmatrix.data* are used as inputs for code *evl.cc*.

```
/*evl.cc*/

#include <math.h>
#include <iostream>
```

```
#include <fstream>
#include <sstream>
#include "nrutil.h"
#include <stdio.h>
#include <stdlib.h>

using namespace std;

extern void tqli (double d[], double e[], int n, double **ev);

double *e;
double *d;
double **ev;
int n,m,l,j,iter,i,k;
double pom,s,r,p,g,f,dd,c,b,dmin;
char sused[500];
main (int argc, char **argv)
{

n=atol(argv[1]);

FILE *h,*h1;
h=fopen("eigenvalue","a");
h1=fopen("eigenvector","a");

ev=new double*[n+1];
d=new double[n+1];
e=new double[n+1];
for(i=0;i<=n;i++) ev[i]=new double[n+1];
i=1;
ifstream file ("diag");
if(!file.is_open()) cout << "file se ne moze otvoriti\n"; /*input data from
file*/
else
   {
   while(!file.eof())
     {
     file.getline(sused,100, '\n');
     pom=atof(sused);
     d[i]=pom;
     i++;
```

```
     }
   }
ifstream file1 ("subdiag");
i=1;
if(!file1.is_open()) cout << "file can not be oppend\n"; /*input data from
file*/
else
   {
   while(!file1.eof())
     {
     file1.getline(sused,100, '\n');
     pom=atof(sused);
     e[i]=pom;
     i++;
     }
   }
ifstream file2 ("tdmatrica.txt");
if(!file2.is_open()) cout << "file can not be oppend\n"; /*input data from
file*/
else
   {
   i=1;
   j=1;
   while(!file2.eof())
    {
    file2.getline(sused,500, ' ');
    pom=atof(sused);
    if(pom==0.00000000 && j==n+1)
    {
    i++;
    j=1;
    }
    else
    {
    ev[i][j]=pom;
    j++;
    }
   }
   }

tqli(d,e,n,ev);
```

```
dmin=d[1];
for(i=1;i<=n;i++)
  {
  fprintf(h,"%.16f\n",d[i]);
  if(!(i==1) && d[i]>0 && d[i]<dmin) dmin=d[i];
  for(l=1;l<=n;l++)
    {
    fprintf(h1,"%.16f ",ev[i][l]);
    }
  fprintf(h1,"0 \n");
  }

for(i=1;i<n+1;i++) delete [] ev[i];
delete [] ev;
delete [] d;
delete [] e;
return EXIT_SUCCESS;
}
```

The output file *eigenvalues* contains a list of eigenvalue while *eigenvectors* contains the matrix of eigenvectors (i-th column corresponds is eigenvector of i-th eigenvalue).

Appendix B

Data collection

The data from B92 and BBC blog were collected using three different *python* scripts:

- *collect_posts.py*
- *collect_users.py*
- *collect_text.py*

The global structure of the scripts is the same for both Blogs, with some differences which depend on *html* code on the specific Web site. Hear we give scripts for BBC Blog.
The script *collect_posts.py* is used for post crawling. The input file is the list of *http* addresses of Blog pages (*list_blog pages*), every Blog page belongs to one author, and the outputs are written into two files: the list of *http* addresses of all post (*list_post_address*), and the file which contains post IDs together with author IDs and posting times (*list_posts*). The *html* Blog pages is loaded as a string and specific phrases related to specific term (*http* address, post IDs, author IDs and posting time) are searched trough the string.

```
#!/usr/bin/python
g=open("list_posts","w");
g1=open("list_post_adress","w")
f=open("list_blog_pages","r")
import urllib
for line in f.readlines():
  p=line.find(" ")
  n=len(line)
  url=line[0:p]
  print url
  sif=line[p+1:n-1]
```

```
f=urllib.urlopen(url);
s=f.read();
i=0;
while((s[i:].find('" id="entry-'))!=-1):
  idx=s[i:].find('" id="entry-')
  idx1=s[i+idx:].find('">')
  psid=sif+s[i+idx+12:i+idx+idx1]
  i=i+idx+idx1
  asidx=s[i:].find('<a href="')
  asidx1=s[i+asidx:].find('html')
  address=s[i+asidx+9:i+asidx+asidx1+4]
  i=i+asidx+asidx1+4
  tiidx=s[i:].find('">')
  tiidx1=s[i+tiidx:].find('</a')
  title=s[i+tiidx+2:i+tiidx+tiidx1]
  i=i+tiidx+tiidx1
  aidx=s[i:].find('"vcard author">')
  aidx1=s[i+aidx+17:].find('">')
  aidx2=s[i+aidx+17+aidx1+2:].find("</a>")
  ss=s[i+aidx+17+aidx1+2:i+aidx+aidx1+19+aidx2]
  i=i+aidx+aidx1+19+aidx2
  ss1=ss.replace(' ','_')
  didx=s[i:].find('title="')
  date=s[i+didx+7:i+didx+7+16]
  i=i+didx+23
  g.write(ss1)
  g.write(" ")
  g.write(title)
  g.write(" ")
  g.write(psid)
  g.write(" ")
  g.write(date)
  g.write("\n")
  g1.write(address)
  g1.write(" ")
  g1.write(psid)
  g1.write(" ")
  g1.write(date)
  g1.write(" ")
  g1.write(ss1)
  g1.write("\n")
```

The *list_post$_a$ddess* is used as input for the second script, *collect_users.py*, used for user and comment crawling. The output file *posts_comments_list* contains the following information: author ID, posts/comment ID, posting time, number of comments, number of recommendations, the list of users who left comment on posts. In the case of BBC Blog the option to comment a comment is not available and the lists of users is left empty. The script in the case of B92 Blog has to be changed for this option to be taken into account.

```
#!/usr/bin/python
g=open("posts_comments_list","w");
f=open("list_post_adress","r")
import urllib
for line in f.readlines():
  p=line.find(" ")
  n=len(line)
  p1=line[p+1:].find(" ")
  p2=line[p+p1+2:].find(" ")
  url=line[0:p]
  print url
  psid=line[p+1:p1+p+1]
  psid1='ps'+psid
  psdate=line[p+p1+2:p+p1+2+p2]
  pauid=line[p+p1+p2+3:n-1]
  f=urllib.urlopen(url);
  s=f.read();
  i=0;
  br=0
  uslist=''
  while((s[i:].find('comment-number">'))!=-1):
    idx=s[i:].find('comment-number">')
    idx1=s[i+idx:].find('. </span>')
    num=s[i+idx+16:i+idx+idx1]
    cmid=num+'cm'+psid
    i=i+idx+idx1
    tidx=s[i:].find("time")
    tidx1=s[i+tidx:].find('">')
    tidx2=s[i+tidx:].find("</a>")
    cmtime=s[i+tidx+tidx1+2:i+tidx+tidx2]
    i=i+tidx+tidx2
    didx=s[i:].find('"date">')
    didx1=s[i+didx:].find("</span>")
    cmdate=s[i+didx+7:i+didx+didx1]
```

```
    cmdate1=cmdate.replace(' ','_')
    i=i+didx+didx1
    uidx=s[i:].find('userid=')
    uidx1=s[i+uidx:].find('">')
    usid='us'+s[i+uidx+7:i+uidx+uidx1]
    i=i+uidx+uidx1
    br=br+1
    g.write(usid)
    g.write(" ")
    g.write(cmid)
    g.write(" ")
    g.write(cmdate1)
    g.write("T")
    g.write(cmtime)
    g.write(" 0 0 EndUsers\n")
    uslist=uslist+usid+' '
  g.write(pauid)
  g.write(" ")
  g.write(psid1)
  g.write(" ")
  g.write(psdate)
  g.write(" ")
  g.write(str(br))
  g.write(" 0 ")
  g.write(uslist)
  g.write("\n")
```

The third script, *collect_text.py* is used only for BBC Blog, where the text of the posts and comments is written in English and can be classified. The input file for this script is again the list of post's addresses, while output file $posts_comments_text$ contain the IDs of comments and their cleaned text.

```
#!/usr/bin/python
g=open("posts_comments_text","w");
f=open("list_post_adress","r")
sc=''
prazno='prazno'
import urllib
for line in f.readlines():
  p=line.find(" ")
  n=len(line)
  p1=line[p+1:].find(" ")
```

```
p2=line[p+p1+2:].find(" ")
url=line[0:p]
#print url
psid=line[p+1:p1+p+1]
psid1='ps'+psid
psdate=line[p+p1+2:p+p1+2+p2]
pauid=line[p+p1+p2+3:n-1]
f=urllib.urlopen(url);
s=f.read();
i=0;
br=0
pcindex=s[i:].find("post_content");
pcindex1=s[i+pcindex:].find("<p>");
pcindex2=s[i+pcindex:].find("<br />")
pctxt=s[i+pcindex+pcindex1:i+pcindex+pcindex2]
pctxt=pctxt.replace('<p>','');
pctxt=pctxt.replace('</p>','');
pctxt=pctxt.replace("\n",'');
pctxt=pctxt.replace('.','. ')
pctxt=pctxt.replace('"','')
pctxt=pctxt.replace("'","")
pctxt=pctxt.replace('<a href','<br ')
pctxt=pctxt.replace('/a>','/> ')
pctxt=pctxt.replace('<strong','<br')
pctxt=pctxt.replace('/strong>','/>')
pctxt=pctxt.replace('^M','')
while (pctxt.count('http')!=0):
  rpidx=pctxt.find('http')
  rpidx1=pctxt[0:rpidx].find('<')
  rpidx2=pctxt[rpidx1:].find('/>')
  pcstr=pctxt[rpidx1:rpidx1+rpidx2+2]
  pctxt=pctxt.replace(pcstr,'',1)
i=i+pcindex+pcindex2
uslist=''
while((s[i:].find('comment-number">'))!=-1):
  idx=s[i:].find('comment-number">')
  idx1=s[i+idx:].find('. </span>')
  num=s[i+idx+16:i+idx+idx1]
  cmid=num+'cm'+psid
  i=i+idx+idx1
  tidx=s[i:].find("time")
```

```
tidx1=s[i+tidx:].find('">')
tidx2=s[i+tidx:].find("</a>")
cmtime=s[i+tidx+tidx1+2:i+tidx+tidx2]
i=i+tidx+tidx2
didx=s[i:].find('"date">')
didx1=s[i+didx:].find("</span>")
cmdate=s[i+didx+7:i+didx+didx1]
cmdate1=cmdate.replace(' ','_')
i=i+didx+didx1
uidx=s[i:].find('userid=')
uidx1=s[i+uidx:].find('" ')
usid='us'+s[i+uidx+7:i+uidx+uidx1]
i=i+uidx+uidx1
cmindex=s[i:].find('comment-text');
cmindex1=s[i+cmindex+12:].find('">');
pcmtxt=s[i+cmindex+12:i+cmindex+12+cmindex1]
ss='-moderation'
if pcmtxt==ss:
  cmtxt=''
else:
  cmindex2=s[i+cmindex+14+cmindex1:].find('</p>');
  cmtxt=s[i+cmindex+14+cmindex1:i+cmindex+14+cmindex1+cmindex2]
i=i+cmindex+14+cmindex1+cmindex2
cmtxt=cmtxt.replace('\n','')
cmtxt=cmtxt.replace('"','')
cmtxt=cmtxt.replace("'","")
cmtxt=cmtxt.replace('<a href','<br ')
cmtxt=cmtxt.replace('/a>','/> ')
cmtxt=cmtxt.replace('<strong','<br')
cmtxt=cmtxt.replace('/strong>','/>')
cmtxt=cmtxt.replace('^M','')
while (cmtxt.count('http')!=0):
  rpidx=cmtxt.find('http')
  rpidx1=cmtxt[0:rpidx].find('<')
  rpidx2=cmtxt[rpidx1:].find('/>')
  cmstr=cmtxt[rpidx1:rpidx1+rpidx2+2]
  cmtxt=cmtxt.replace(cmstr,'',1)
br=br+1
if cmtxt==sc:
  print prazno
else:
```

```
    g.write(usid)
    g.write("|")
    g.write(cmid)
    g.write("|")
    g.write(cmdate1)
    g.write("T")
    g.write(cmtime)
    g.write("|0|")
    g.write(cmtxt)
    g.write("\n")
if pctxt==sc:
  print prazno
else:
  g.write(pauid)
  g.write("|")
  g.write(psid1)
  g.write("|")
  g.write(psdate)
  g.write("|")
  g.write(str(br))
  g.write("|")
  g.write(pctxt)
  g.write("\n")
```

List of publications used for this thesis

1. M.Mitrović, B. Tadić, "Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities", *Phys. Rev., E*, vol. **80**(2), str. 026123-1-026123-12, (2009).

2. M. Mitrović, B. Tadić, "Bloggers behavior and emergent communities in Blog space", *Eur. Phys. J. B*, vol. **73** (2), (2010).

3. M.Mitrović, G.Paltoglou and B.Tadić, "Networks and emotion-driven user communities at popular blogs", *Eur. Phys. J. B* vol. **77**(4), pp. 597-609, (2010).

4. M.Mitrović, G. Paltoglou and B. Tadić, "Quantitative analysis of bloggers' collective behavior powered by emotions", *JSTAT* 2, P02005-1-P02001-16, (2011).

Bibliography

[1] J. Kleinberg. The Convergence of Social and technological Networks. *Communications of the ACM*, 51:66–72, 2008.

[2] J. Giles. Social science lines up its biggest challenges. *Nature*, 470:18–19, 2011.

[3] A. Cho. Ourselves and Our Interactions: The Ultimate Physics Problem? *Science*, 325(5939):406–408, 2009.

[4] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009.

[5] F. Skopik, D. Schall, H. Psaier, M. Treiber, and S. Dustdar. Towards social crowd environments using service-oriented architectures. *Information Technology*, 3:108–116, 2011.

[6] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the Web graph. *J. Phys. A: Math. Theor.*, 41:224017, 2008.

[7] B. Tadić. Dynamics of directed graphs: the world-wide Web. *Physica A: Statistical Mechanics and its Applications*, 293:273–284, 2001.

[8] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E*, 73(3):036127, 2006.

[9] R. D. Malmgren, D. B. Stouffer, A. S. L. O. Campanharo, and L. A. Amaral. On Universality in Human Correspondence Activity. *Science*, 325(5948):1696–1700, 2009.

[10] P.S. Dodds and C.M. Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, 2009.

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

[12] M. Mitrović and B. Tadić. Bloggers Behavior and Emergent Communities in Blog Space. *Eur. Phys. J. B*, 73(2):293–301, 2010.

[13] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks. *Proceedings of the National Academy of Sciences USA*, 107(31):13636–13641, 2010.

[14] S. Gonzalez-Bailon, R.E. Banchs, and A. Kaltenbrunner. Emotional Reactions and the Pulse of Public Opinion: Measuring the Impact of Political Events on the Sentiment of Online Discussions, 2010. e-print arXiv:1009.4019.

[15] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61:2544–2558, 2010.

[16] M. Mitrović, G. Paltoglou, and B. Tadić. Networks and emotion-driven user communities at popular blogs. *Eur. Phys. J. B*, 77:597–609, 2010.

[17] M. Mitrović and B. Tadić. *Emergence and structure of cybercommunities*. Springer, Berlin, 2011.

[18] M. Mitrović, G. Paltoglou, and B. Tadić. Quantitative analysis of bloggers' collective behavior powered by emotions.

Journal of Statistical Mechanics: Theory and Experiment, 2011(02):P02005+, 2011.

[19] P.S. Dodds, K.D. Harris, I.M. Koloumann, C.A. Bliss, and C.M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometric and Twitter. *arXiv:1101.5120v3*, 2011.

[20] F. Fu, L. Liu, K. Yang, and L. Wang. The structure of self-organized blogosphere. *arXiv:0607361*, 2006.

[21] Y. Sano and M. Takayasu. Macroscopic and microscopic statistical properties observed in blog entries. *arXiv:0906.1744*, 2009.

[22] E. Hatfield, J.T Cacioppo, and R.L. Rapson. *Emotional Contagion (Studies in Emotion and Social Interaction*. Cambridge University Press, Cambridge, UK, 1994.

[23] R. W. Larson and D. M. Almeida. Emotional transmission in the daily lives of families: a new paradigm for studying family process. *J. Marriage Fam.*, 61:5–20, 2010.

[24] M. J. Howes, J. E. Hokanson, and D. A. Loewenstein. Induction of depressive affect after prolonged exposure to a mildly depressed individual. *J. Personal. Soc. Psychol.*, 49:1110–1113, 1985.

[25] S. G. Barsade. The ripple effect: emotional contagion and its influence on group behavior. *Admin. Sci. Q.*, 47:644–675, 2002.

[26] P. Ekman. *Emotion in the human face*. Cambridge University Press, Cambridge, UK, 1982.

[27] S.R. Fussell. *The verbal communication of emotion*. Lawrence Erlbaum Associates, Mahwan, NJ, 2002.

[28] J.B. Walther. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research*, 19:52–60, 1992.

[29] J.B. Walther, T. Loh, and L. Granka. Let me count the ways: The interchange of verbal and nonverbal cues in computermediated and face-to-face affinity. *Journal of Language and Social Psychology*, 24:36–65, 2005.

[30] M. Thelwall, A. Byrne, and M. Goody. Which types of news story attract bloggers? *Informetionresearch*, 12:327+–22, 2007.

[31] M. Thelwall, D. Wilkinson, and S. Uppal. Data Mining Emotion in Social Network Communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, 2009.

[32] Jukka-Pekka Onnela and Felix Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, 2010.

[33] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[34] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, 2008.

[35] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[36] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[37] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[38] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586):1551–1555, 2002.

[39] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.

[40] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[41] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[42] B. Tadić, G. J. Rodgers, and S. Thurner. Transport on Complex Networks: Flow, Jamming and Optimization. *International Journal of Bifurcation and Chaos*, 17(7):2363–2385, 2007.

[43] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications (ASA-SIAM Series on Statistics and Applied Probability)*. SIAM, Society for Industrial and Applied Mathematics, 2007.

[44] L. Danon, A. Díaz-Guilera, and A. Arenas. The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 11:P11010, 2006.

[45] S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. *Phys. Rev. E*, 70:056104, Nov 2004.

[46] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[47] B. Krishnamurthy and J. Wang. On network-aware clustering of Web clients. *SIGCOMM Comput. Commun. Rev.*, 30:97–110, 2000.

[48] P. K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao. A Graph Based Approach to Extract a Neighborhood Customer Community for Collaborative Filtering. In *Proceedings of the Second International Workshop on Databases in Networked Information Systems*, DNIS '02, pages 188–200, London, UK, UK, 2002. Springer-Verlag.

[49] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[50] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569, 2007.

[51] M. Mitrović and B. Tadić. Search of Weighted Subgraphs on Complex Networks with Maximum Likelihood Methods. *Lecture Notes in Computer Science*, 5102:551–558, 2008.

[52] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Sciences (Cambridge Nonlinear Science Series)*. Cambridge University Press, 2003.

[53] A. Arenas, A. Diaz-Guilera, J. Kurths, Y. Moreno, and CS Zhou. Synchronization in complex networks. *Physics Reports*, 469:93–153, 2008.

[54] Y. Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer–Verlag, New York, 1984.

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

[56] V. Darley. Emergent Phenomena and Complexity. In *Artificial Life IV. Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pages 411–416. MIT Press, 1994.

[57] S. Wolfram. Cellular automata as models of complexity. *Nature*, 311:419–424, 1984.

[58] M. Gardner. The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223:120–123, 1970.

[59] B. Chopard and M. Droz. *Cellular Automata Modeling of Physical Systems*. Cambridge University Press, 1998.

[60] B. G. Ermentrout and L. Edelstein-Keshet. Cellular automata approaches to biological modeling. *Journal of theoretical biology*, 160:97–133, 1993.

[61] J. Hardy, O. de Pazzis, and Y. Pomeau. Molecular dynamics of a classical lattice gas: Transport properties and time correlation functions. *Phys. Rev. A*, 13:1949–1961, 1976.

[62] Sauro Succi, Roberto Benzi, and Francisco Higuera. The lattice boltzmann equation: A new tool for computational fluid-dynamics. *Physica D: Nonlinear Phenomena*, 47(1-2):219 – 230, 1991.

[63] R. Capuccio, G. Cattaneo, D. Ciucci, and U. Jocher. Parallel implementation of a cellular automat based model for coffee percolation. *Parallel Computing*, 2000.

[64] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the $1/f$ noise. *Phys. Rev. Lett.*, 59:381–384, 1987.

[65] H. J. JensenD. *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems (Cambridge Lecture Notes in Physics)*. Cambridge University Press, 1998.

[66] D. Dhar. Self-organized critical state of sandpile automaton models. *Phys. Rev. Lett.*, 64(14):1613–1616, 1990.

[67] D. Dhar. Theoretical studies of self-organized criticality. *Physica A: Statistical Mechanics and its Applications*, 369(1):29–70, 2006.

[68] A. Kirchner and A. Schadschneider. Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. *Physica A: Statistical Mechanics and its Applications*, 312(1-2):260–276, 2002.

[69] F. Bagnoli, A. Berrones, and F. Franci. De gustibus disputandum (forecasting opinions by knowledge networks). *Physica A: Statistical Mechanics and its Applications*, 332:509–518, 2004.

[70] M. Bartolozzi and A. W. Thomas. Stochastic cellular automata model for stock market dynamics. *Phys. Rev. E*, 69:046112, 2004.

[71] G. N. Gilbert. *Agent-based models*. Sage Publications, Los Angeles, 2008.

[72] J. M. Epstein and R. L. Axtell. *Growing Artificial Societies: Social Science from the Bottom Up (Complex Adaptive Systems)*. The MIT Press, 1996.

[73] R. Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, 1997.

[74] C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, volume 21, pages 25–34, New York, NY, USA, 1987. ACM.

[75] Edwin Hutchins, Brian Hazlehurst, and R. Conte (eds. How To Invent A Lexicon: The Development Of Shared Symbols In Interaction. In *Artificial Societies: The Computer Simulation of Social Life*, pages 157–189. UCL Press, 1995.

[76] P. Terna. A laboratory for agent based computational economics. In *Simulating Social Phenomena, Lecture Notes in Economics and Mathematical Systems*. Berlin: Springer-Verlag, 1997.

[77] E. Chattoe and N. Gilbert. A Simulation of Adaptation Mechanisms in Budgetary Decision Making. In *Simulating Social Phenomena, Lecture Notes in Economics and Mathematical Systems*, volume 456, pages 401–418. Berlin: Springer-Verlag, 1997.

[78] A. L. C. Ana, J. Wahle, and F. Klugl. Agents in Traffic Modelling - From Reactive to Social Behaviour. *Lecture Notes in Artificial Intelligence*, 1701:303–306, 1999.

[79] T. Lux and M. Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500, 1999.

[80] S. N. Dorogovtsev. *Lectures on complex networks*. Oxford University Press, 2010.

[81] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, New York, NY, USA, 2004.

[82] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and Correlation Properties of the Internet. *Phys. Rev. Lett.*, 87:258701, 2001.

[83] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.

[84] J. Grujić, M. Mitrović, and B. Tadić. Mixing patterns and communities on bipartite graphs on web-based social interactions.

In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–8, 2009.

[85] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, 2002.

[86] L. Donetti and M. A. Muñoz. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10):P10012, 2004.

[87] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2001.

[88] S. Maletić, M. Rajković, and D. Vasiljević. Simplicial Complexes of Networks and Their Statistical Properties. In *ICCS (2)*, volume 5102 of *Lecture Notes in Computer Science*, pages 568–575. Springer, 2008.

[89] P. N. McGraw and M. Menzinger. Laplacian spectra as a diagnostic tool for network structure and dynamics. *Phys. Rev. E*, 77(3):031102–+, 2008.

[90] J.A. Almendral and A. Daz-Guilera. Dynamical and spectral properties of complex networks. *New Journal of Physics*, 9(6):187, 2007.

[91] D.C. Bell, J.S. Atkinson, and J.W. Carlson. Centrality measures for disease transmission networks. *Social Networks*, 21(1):1–21, 1999.

[92] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of Growing Networks with Preferential Linking. *Phys. Rev. Lett.*, 85:4633–4636, 2000.

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

[94] B. Mandelbrot and R. L. Hudson. *The Misbehavior of Markets: A Fractal View of Financial Turbulence.* Basic Books, 2006.

[95] R. Crane, F. Schweitzer, and D. Sornette. Power law signature of media exposure in human response waiting time distributions. *Phys. Rev. E*, 81(5):056101, 2010.

[96] G. Grinstein and R. Linsker. Power-law and exponential tails in a stochastic priority-based model queue. *Phys. Rev. E*, 77(1):012101–+, 2008.

[97] B. Tadić. Temporal fractal structures: origin of power laws in the world-wide Web. *Physica A: Statistical Mechanics and its Applications*, 314(1-4):278–283, 2002.

[98] T. Zhou, L.-L. Jiang, R.-Q. Su, and Y.-C. Zhang. Effect of initial configuration on network-based recommendation. *Eyrophys. Lett.*, 81:58004, 2008.

[99] B. Tadić, S. Thurner, and G. J. Rodgers. Traffic on complex networks: Towards understanding global statistical properties from microscopic density fluctuations. *Phys. Rev. E*, 69(3):036102, 2004.

[100] Z. Eisler and J. Kertész. Random walks on complex networks with inhomogeneous impact. *Phys. Rev. E*, 71(5):057104, 2005.

[101] G.J. Janacek. *Practical time series.* Arnold Publishers, 2001.

[102] Á. Corral. Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys. Rev. Lett.*, 92(10):108501, 2004.

[103] D. Sornette, S. Utkin, and A. Saichev. Solution of the nonlinear theory and tests of earthquake recurrence times. *Phys. Rev. E*, 77(6):066109, 2008.

[104] D. Spasojević, S. Bukvić, S. Milošević, and G. Stanley. Barkhausen noise: elementary signals, power laws, and scaling relations. *Phys. Rev. E*, 54:2531, 1996.

[105] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1 edition, 2008.

[106] C. D. Manning and H. Schüetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, 1999.

[107] I.H. Witten and T.C. Bell. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *Information Theory, IEEE Transactions on*, 37(4):1085–1094, 1991.

[108] Macdonald C. and I. Ounis. The TREC Blogs06 Collection : Creating and Analysing a Blog Test Collection. Technical report, Department of Computing Science, University of Glasgow, 2006.

[109] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2008 blog track. In *The Sixteenth Text REtrieval Conference (TREC 2008) Proceedings*, 2008.

[110] C. Saarni. *The development of emotional competence*. Guilford Press, New York, NY, US, 1999.

[111] J. A. Russell and B. Fehr. Fuzzy concepts in a fuzzy hierarchy: varieties of anger. *Journal of Personality and Social Psychology*, 67(2):186–205, 1994.

[112] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[113] Margaret M Bradley and Peter J Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida., 1999.

[114] J. Ahn, S. Gobron, Q. Silvestre, and D. Thalmann. Asymmetrical Facial Expressions based on an Advanced Interpretation of Two-dimensional Russell's Emotional Model. In *pre-print*, 2010.

[115] D. Derks, A. H. Fischer, and A. E. R. Bos. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3), 2008.

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

[117] F. Schweitzer. *Browning Agents and Active Particles: Collective Dynamics in the Natural and Social Sciences.* Springer-Verlag Berlin Heidelberg, 2007.

[118] F. Schweitzer and D. Garcia. An agent-based model of collective emotions in online communities. *Eur. Phys. J. B*, 77(4):533–545, 2010.

[119] L. Feldman Barrett and J. A. Russell. Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, 74(4):967–984, 1998.

[120] K. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[121] L. Buzna, S. Lozano, and A. Díaz-Guilera. Synchronization in symmetric bipolar population networks. *Phys. Rev. E*, 80:066120, 2009.

[122] Z. Levnajić, , and B. Tadić. Self-organization in trees and motifs of two-dimensional chaotic maps with time delay. *Journal of Statistical Physics: Theory and Experiment*, 2008.

[123] M. M. Bradley, M. Codispoti, B. N. Cuthbert, and P. J. Lang. Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion*, 1(3):276–298, 2001.

[124] J. H Fowler and N. A. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *British Medicine Journal*, 337:a2338, 2008.

[125] M. Mitrović and B. Tadić. Patterns of Emotional Blogging and Emergence of Communities: Agent-Based Model on Bipartite Networks. *arXiv:1110.5057*, abs/1110.5057, 2011.

[126] M. Francl. Blogging on the sidelines: Bloggers shouldn't be relegated to the sidelines of the scientific literature. *Nature chemistry*, 3(3):183–184, 2011.

[127] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA, 1992.