QLectives – Socially Intelligent Systems for Quality
Project no. 231200

Instrument: Large-scale integrating project (IP)
Programme: FP7-ICT

Deliverable D.2.2.1

Algorithms for detecting, emerging, and sustaining cooperative community structures

Submission date: 2013-03-01

Start date of project: 2009-03-01 Duration: 48 months

Organisation name of lead contractor for this deliverable:
University of Fribourg

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidental, only for members of the consortium (including the Commission Services)	

Document information

1.1 Author(s)

Author	Organisation	E-mail
Matúš Medo	University of Fribourg	matus.medo@unifr.ch

1.2 Other contributors

Name	Organisation	E-mail
Róbert Ormándi	University of Szeged	ormandi@inf.u-szeged.hu
István Hegedűs	University of Szeged	ihegedus@inf.u-szeged.hu
Márk Jelasity	University of Szeged	jelasity@inf.u-szeged.hu
Sergi Lozano	ETH Zurich	slozano@ethz.ch

1.3 Document history

Version#	Date	Change
V1.0	17 January, 2013	First draft
V1.1	25 January, 2013	Approved version to be submitted to EU (small corrections)

1.4 Document data

Keywords	complex networks, distributed algorithms
Editor address data	matus.medo@unifr.ch
Delivery date	01 03, 2013

1.5 Distribution list

Date	Issue	E-mail
	Consortium members	QLECTIVES@list.surrey.ac.uk
	Project officer	Roumen.BORISSOV@ec.europa.eu
	EC archive	INFSO-ICT-231200@ec.europa.eu

QLectives Consortium

This document is part of a research project funded by the ICT Programme of the Commission of the European Communities as grant number ICT-2009-231200.

University of Surrey (Coordinator)
Department of Sociology/Centre
for Research in Social Simulation
Guildford GU2 7XH
Surrey
United Kingdom
Contact person: Prof. Nigel Gilbert
E-mail: n.gilbert@surrey.ac.uk

University of Fribourg
Department of Physics
Fribourg 1700
Switzerland
Contact person: Prof. Yi-Cheng Zhang
E-mail: yi-cheng.zhang@unifr.ch

Technical University of Delft
Department of Software Technology
Delft, 2628 CN
Netherlands
Contact Person: Dr Johan Pouwelse
E-mail: j.a.pouwelse@tudelft.nl

University of Warsaw
Faculty of Psychology
Warsaw 00927
Poland
Contact Person: Prof. Andrzej Nowak
E-mail: nowak@fau.edu

ETH Zurich
Chair of Sociology, in particular
Modelling and Simulation
Zurich, CH-8092
Switzerland
Contact person: Prof. Dirk Helbing
E-mail: dhelbing@ethz.ch

Centre National de la Recherche Scientifique, CNRS
Paris 75006,
France
Contact person: Dr. Camille ROTH
E-mail: camille.roth@polytechnique.edu

University of Szeged
MTA-SZTE Research Group on
Artificial Intelligence
Szeged 6720, Hungary
Contact person: Dr Mark Jelasity
E-mail: jelasity@inf.u-szeged.hu

Institut für Rundfunktechnik GmbH
Munich 80939
Germany
Contact person: Dr. Christoph Dosch
E-mail: dosch@irt.de

QLectives introduction

QLectives is a project bringing together top social modelers, peer-to-peer engineers and physicists to design and deploy next generation self-organising socially intelligent information systems. The project aims to combine three recent trends within information systems:

- **Social networks** - in which people link to others over the Internet to gain value and facilitate collaboration
- **Peer production** - in which people collectively produce informational products and experiences without traditional hierarchies or market incentives
- **Peer-to-Peer systems** - in which software clients running on user machines distribute media and other information without a central server or administrative control

QLectives aims to bring these together to form Quality Collectives, i.e. functional decentralised communities that self-organise and self-maintain for the benefit of the people who comprise them. We aim to generate theory at the social level, design algorithms and deploy prototypes targeted towards two application domains:

- **QMedia** - an interactive peer-to-peer media distribution system (including live streaming), providing fully distributed social filtering and recommendation for quality
- **QScience** - a distributed platform for scientists allowing them to locate or form new communities and quality reviewing mechanisms, which are transparent and promote

The approach of the QLectives project is unique in that it brings together a highly inter-disciplinary team applied to specific real world problems. The project applies a scientific approach to research by formulating theories, applying them to real systems and then performing detailed measurements of system and user behaviour to validate or modify our theories if necessary. The two applications will be based on two existing user communities comprising several thousand people – so-called "Living labs", media sharing community tribler.org; and the scientific collaboration forum EconoPhysics.

Executive summary

In this deliverable, we start by briefly reviewing relevant work by QLectives partners which has been presented in other deliverables so we do not fully expound it here. We continue with two chapters on fundamental issues.

Chapter 3 addresses data in real world recommender applications which often feature fat-tailed distributions of the number of times individual items have been rated or favored. We propose a model to simulate such data. The model is mainly based on social interactions and opinion formation taking place on a complex network with a given topology. A threshold mechanism is used to govern the decision making process that determines whether a user is or is not interested in an item. We demonstrate the validity of the model by fitting attendance distributions from different real data sets. The model is mathematically analyzed by investigating its master equation. Our approach provides an attempt to understand recommender system's data as a social process. The model can serve as a starting point to generate artificial data sets useful for testing and evaluating recommender systems.

Chapter 4 studies behavior of populations under stress when conflicting scenarios often give rise to social cohesion in human groups (collectives). By means of a simple computational model, we explore a dynamic perspective of social cohesion in populations under stress. Dynamics are driven by the co-evolution of structural and cognitive dimensions. Submitted to sudden variations on its environmental conflict level, the model is able to reproduce certain characteristics previously observed in real populations in situations of emergency or crisis. A closer analysis of the results, observing both structural and cognitive together, uncovers a causal path from the level of conflict suffered by a population to variations on its social cohesiveness.

Chapter 5 notes a typical feature of real datasets used for benchmarking of distributed recommendation algorithms and proposes algorithms which address it. Offering personalized recommendation as a service in fully distributed applications such as file-sharing, distributed search, social networking, P2P television, etc, is an increasingly important problem. In such networked environments recommender algorithms should meet the same performance and reliability requirements as in centralized services. To achieve this is a challenge because a large amount of distributed data needs to be managed, and at the same time additional constraints need to be taken into account such as balancing resource usage over the network. In this paper we focus on a common component of many fully distributed recommender systems, namely the overlay network. We point out that the overlay topologies that are typically defined by node similarity have highly unbalanced degree distributions in a wide range of available benchmark datasets: a fact that has important—but so far largely overlooked—consequences on the load balancing of overlay protocols. We propose algorithms with

a favorable convergence speed and prediction accuracy that also take load balancing into account. We perform extensive simulation experiments with the proposed algorithms, and compare them with known algorithms from related work on well-known benchmark datasets. The basic idea is general in the sense it is independent from the applied similarity metric used in the recommendation algorithm. As a result, this approach can be applied to more sophisticated metrics, even those learned by machine-learning techniques. This type of learning can be performed in fully distributed systems by applying the mechanisms proposed by the Szeged team in [88, 89].

In summary, this deliverable describes how social interactions and collectives influence and overlap with opinion formation and how the resulting informational structures (in this case, fat-tailed degree distributions in overlay networks) need to be taken into account in the design of distributed algorithms.

Contents

Chapter 1

Introduction

In this deliverable, we start by briefly reviewing relevant work by QLectives partners which has been presented in other deliverables so we do not fully expound it here. We continue with two chapters on fundamental issues. In Chapter 3, we show that social interactions and opinion formation on a complex network can give rise to fat-tailed popularity distribution of content. This work provides an attempt to understand recommender system's data as a social process. It can also serve as a starting point to generate artificial data sets useful for testing and evaluating of recommender and reputation systems. Work was presented at workshop Decisions@RecSys 2012 in Dublin [19]. In Chapter 4, we use a computational model to explore a dynamic perspective of social cohesion in populations under stress where dynamics is driven by the co-evolution of structural and cognitive dimensions. The model is able to reproduce certain characteristics previously observed in real populations in situations of emergency or crisis. This chapter is based on [69]. In Chapter 5, we directly address the third objective of this deliverable: "Provide a basis for efficiently distributed implementation of the proposed algorithms". We focus on a common component of many fully distributed recommender systems, namely the overlay network. Overlay topologies are typically defined by node similarity and have highly unbalanced degree distributions in a wide range of available benchmark datasets. This has profound consequences on the load balancing of overlay protocols. We propose algorithms with a favorable convergence speed and prediction accuracy that also take load balancing into account. While the proposed mechanism is specifically designed to fully distributed (P2P) environments where numerous constraints must be satisfied, the basic idea can be applied successively in other systems with similar requirements. This chapter is based on [87].

Chapter 2

Related work by QLectives partners

2.1 Algorithm for quality, trust and reputation in online communities

In the Internet era the information overload and the challenge to detect quality content has raised the issue of how to rank both resources and users in online communities. We proposed a novel and generalized ranking algorithm for bipartite systems to assign quality values to objects and reputation values to users. This method, which we named QTR (Quality, Trust and Reputation), also exploits the information coming from the users' social relationships. QTR is a generalized algorithm in the sense that it can be easily adapted to different situations (*e.g.*,by giving more weight to certain kind of actions, or to a particular behavior of users). We tested our method on two different datasets, the EconoPhysics forum online community and the Last.fm online radio and social network, which are particularly suited for our generalized algorithm because they feature various levels of interactions (uploading a paper, downloading a paper, viewing a paper's abstract in the case of the EconoPhysics Forum) and multi-level network where the bipartite user-item structure is superimposed on a social network of users (in the case of the Last.fm data). The results of our study are twofold. We first confirmed that ranking is a difficult task, and that an improper algorithm or a peculiarly-structured dataset can lead to biased results. To address this, we proposed a form of the QTR which is efficient in avoiding such bias. In addition, we showed that social relationships can play a valuable role in improving the quality of the resulting rankings. Details about the algorithm and its validation can be found in Deliverable 3.3.1 (Report on model validation and synthesis) and [67].

2.2 Trend prediction in temporal bipartite networks

Online systems where users purchase or collect items of some kind can be effectively represented by temporal bipartite networks where both nodes and links are added with time. We used this representation to predict which items might become popular in the near future. Various prediction methods were evaluated on three distinct datasets originating from popular online services (Movielens, Netflix, and Digg) with weighted popularity increase, where users are weighted by their overall activity, pro-

duced the most promising results. We also showed that the prediction performance can be further enhanced if the social user network is known and contribution of individual users are weighted by their centrality. For more details see Deliverable 1.1.2 (Book manuscript on agent-based models of complex techno-social systems).

Chapter 3

Recommendation systems in the scope of opinion formation: a model

3.1 Introduction

This is the information age. We are witnessing information production and consumption in a speed never seen before. The WEB2.0 paradigm enables consumers and producers to exchange data in a collaborative way benefiting both parties. However, one of the key challenges in our digitally-driven society is information overload [13]. We have the 'pain of choice'. Recommendation systems represent a possible solution to this problem. They have emerged as a research area on its own in the 90s [22, 40, 43, 63, 99]. The interest in recommendation systems increased steadily in recent years, and attracted researchers from different fields [100]. The success of highly rated Internet sites as Amazon, Netflix, YouTube, Yahoo, Last.fm and others is to a large extent based on their recommender engines. Corresponding applications recommend everything from CD/DVD's, movies, jokes, books, web sites to more complex items such as financial services.

The most popular techniques related to recommendation systems are collaborative filtering [15, 22, 43, 51, 60, 63, 97, 102] and content-based filtering [11, 27, 68, 76, 91]. In addition, researchers developed alternative methods inspired by fields as diverse as machine learning, graph theory, and physics [18, 33, 34, 78, 112, 115–117]. Furthermore, recommendation systems have been investigated in connection with trust [5, 73, 74, 85, 110] and personalized web search [17, 23, 104], which constitutes the new research frontier in search engines.

However, there are still many open challenges in the research field of recommendation systems [2, 30, 37, 48, 51, 55, 100]. One key question is connected to the understanding of the user rating mechanism. We build on a well documented influence of social interactions with peers on the decision to vote, favor, or even purchase an item [61, 101]. We propose a model inspired by opinion formation taking place on a complex network with a predefined topology. Our model is able to generate data observed in real world recommender systems. Despite its simplicity, the model is flexible enough to generate a wide range of different patterns. We mathematically analyze the model using a mean field approach to the full Master Equation. Our approach provides an understanding of the data in recommender systems as a product of social

processes. The model can serve as a data generator which is valuable for testing and evaluation purposes for recommender systems.

The rest of this chapter is organized as follows. The model is outlined in Sec. (3.2). Methods, data set descriptions, and validation procedures are in Sec. (3.3). Results are presented in Sec. (3.4). Discussion and an outlook for future research directions are in Sec. (3.5).

3.2 Model

3.2.1 Motivation

Our daily decisions are heavily influenced by various information channels: advertisement, broadcastings, social interactions, and many others. Social ties (word-of-mouth) play a pivotal role in consumers buying decisions [61, 101]. It was demonstrated by many researchers that personal communication and informal information exchange not only influence purchase decisions and opinions, but shape our expectations of a product or service [6, 8, 114]. On the other hand, it was shown [50], that social benefits are a major motivation to participate on opinion platforms. If somebody is influenced by recommendations on an opinion platform like MovieLens or Amazon, social interactions and word-of-mouth in general are additional forces governing the decision making process to purchase or even to rate an object in a particular way [72].

Our model is formulated within an opinion formation framework where social ties play a major role. We shall discuss the following main ingredients of our model:

- Influence-Network (IN)
- Intrinsic-Item-Anticipation (IIA)
- Influence-Dynamics (ID)

Influence Network We call the network where context-relevant information exchange takes place an Influence-Network (IN). Nodes of the IN are people and connections between nodes indicate the influence among them. Note that we put no constraints on the nature of how these connections are realized. They may be purely virtual (over the Internet) or based on physical meetings. We emphasize that INs are domain dependent, i.e., for a given community of users, the Influence Network concerning books may differ greatly (in topology, number of ties, tie strength, etc.) from that concerning another subject such as food or movies. Indeed, one person's opinion leaders (relevant peers) concerning books may be very different from those for food or other subjects. In this scope, we see the INs as domain-restricted views on social networks. It is thus reasonable to assume that Influence Networks are similar to social interaction networks which often exhibit a scale-free topology [12]. However, our model is not restricted to a particular network structure.

Intrinsic-Item-Anticipation Suppose a new product is launched on the market. Advertisement, marketing campaigns, and other efforts to attract customers predate the

launching process and continue after the product started to spread on the market. These efforts influence product-dependent customer anticipation. It is clear that the resulting anticipation is a complex combination of many different components including intrinsic product quality and possibly also suggestions from recommendation systems.

In our model we call the above-described anticipation Intrinsic-Item-Anticipation (IIA) and measure it by a single number. It is based on many external sources, except for the influence generated by social interactions. It is the opinion on something taken by individuals, before they start to discuss the subject with their peers. Furthermore, we assume that an individual will invest resources (time/money) into an object only, if the Intrinsic-Item-Anticipation is above a particular threshold, which we call Critical-Anticipation-Threshold.

Influence-Dynamics The Influence-Dynamics describes how individuals' Intrinsic-Item-Anticipations are altered by information exchange via the connections of the corresponding Influence-Network. From our model's point of view this means the following: an individual's IIA for a particular item i may be shifted due to social interactions with directly connected peers (these interactions thus take place on the corresponding IN), who already experienced the product or service in question. This process can shift the Intrinsic-Item-Anticipation of an individual who did not yet experience product/object i closer to or beyond the critical-anticipation-threshold.

We now summarize the basic ingredients of our model. An individual user's opinions on objects are assembled in two consecutive stages: i) opinion making based on different external sources, including suggestions by recommendation systems and ii) opinion making based on social interactions in the Influence-Network. The second process may shift the opinions generated by the first process.

3.2.2 Mathematical formulation of the model

In this section we firstly describe how individuals' Intrinsic-Item-Anticipations may change due to social interactions taking place on a particular Influence-Network. Secondly, we introduce dynamical processes governing the opinion propagation.

IIA shift We model a possible shift in the IIA as:

$$\hat{f}_{ij} = f_{ij} + \left[\frac{\Theta_j}{k_j}\right]^{(1-\gamma)} . \tag{3.1}$$

where $\hat{f}_{ij}$ is the shifted Intrinsic-Item-Anticipation of individual j for object i, f_{ij} is the unbiased IIA, Θ_j is the number of j's neighbors, who already experienced and liked item i, k_j denotes the total number of j's neighbors in the corresponding IN, and $\gamma \in (0,1)$ quantifies trust of individuals to their peers. An individual j will consume, purchase, or positively rate an item i only if

$$\hat{f}_{ij} \geq \Delta. \tag{3.2}$$

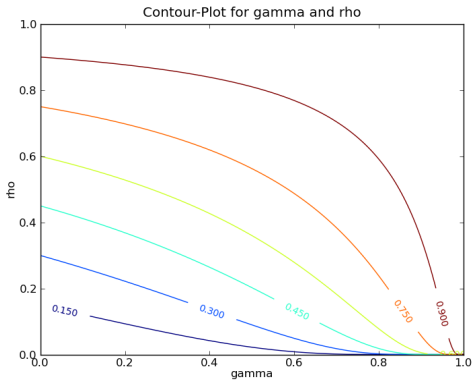

Figure 3.1: Contour plot for γ and $\rho = \Theta_j/k_j$. Numbers inside the plot quantify the shift in the IAA as a function of γ and ρ.

We identify Δ as the Critical-Anticipation-Threshold. Values of f_{ij} are drawn from a probability distribution f_i. Since the IIA for each individual is an aggregate of many different and largely independent contributions, we assume that f_i is normally distributed, $f_i \in \mathcal{N}(\mu_i, \sigma)$. (Unless stated otherwise.) To mimic different item anticipations for different objects i, we draw the mean μ_i from a uniform distribution $U(-\epsilon, \epsilon)$. We maintain μ_i, ϵ, and σ, so that f_i is roughly bounded by $(-1, 1)$, i.e., $-1 \leq \mu - 3\sigma < \mu + 3\sigma \leq 1$. Note that $\hat{f}_{ij}$ can exceed these boundaries after a shift of the corresponding IIA occurs. The second term on the right hand side of Eq. (3.1) is the influence of j's neighborhood weighted by trust γ. To better understand the interplay between γ and the density of attending users in the neighborhood of user i, $\rho := \Theta_j/k_j$, we refer to Fig. 3.1. Trust $\gamma \approx 1$ causes a big shift on the IIA's even for $\rho \approx 0$. On the other hand, $\gamma \approx 0$ needs high ρ to yield a significant IAA shift. These properties are understood as follows: people trusting strongly in their peers need only few positive opinions to be convinced, whereas people trusting less in their social environment need considerable more signals to be influenced.

Influence-Dynamics The Influence-Dynamics proceeds as follows. Firstly, we draw an Influence-Network IN($\mathcal{P}$) with a fixed network topology (power-law, Erdős-Rényi, or another). $\mathcal{P}$ refers to a set of appropriate parameters for the Influence-Network in question (like network type, number of nodes, etc.). The network's topology is not affected by the dynamical processes (opinion propagation) taking place on it. We justify this static scenario by assuming that the time scale of the topology change is much longer then the time scale [1] of opinion spreading in the network. Each node in the Influence-Network corresponds to an individual. For each individual j we draw an unbiased Intrinsic-Item-Anticipation f_{ij} from the predefined probability distribution f_i. At each time step, every individual is in one of the following states: $\{S, A, D\}$. S refers to a susceptible state and corresponds to the initial state for all nodes at $t = 0$. A refers to an attender state and corresponds to an individual with the property $\hat{f}_{ij} \geq \Delta$.

[1] The term time scale denotes a dimensionless quantity and specifies the devisions of time. A shorter time scale means a faster spreading of opinions in the network.

D refers to a denier state with the property $\hat{f}_{ij} < \Delta$ after an information exchange with his/her peers in the Influence-Network happened. An individual in state D or A can not change his/her state anymore. It is clear that an individual in state A cannot back transform to the susceptible state S, since he/she did consume or favor item i and we do not account for multiple attendances in our model. An individual in state D was influenced but not convinced by his opinion leaders (directed connected peers). We make the following assumption here: if individual j's opinion leaders are not able to convince individual j, meaning that individual's j Intrinsic Item Anticipation $\hat{f}_{ij}$ stays below the critical threshold Δ after the influence process, then we assume that j's opinion not to attend object i remains unchanged in the future. Therefore we have the following possible transitions for each node in the influence network: $j_S \rightarrow j_A$ or $j_S \rightarrow j_D$. Node states are updated asynchronously which is more realistic than synchronous updating, especially in social interaction models [24]. The Influence-Dynamics is summarized in Algorithm 1.

Algorithm 1 RecSysMod algorithm. $\mathcal{P}$ contains the configuration parameter for the network. Δ is the Anticipation Threshold and γ denotes the trust. $O \in \mathbb{N}$ is the number of objects to simulate. $G(N, E)$ is the network. N is the set of nodes and E is the set of edges.

procedure RecSysMod_I($\mathcal{P}, \Delta, \gamma, O$)
  $G(N, E) \leftarrow$ GenNetwork($\mathcal{P}$)
  for all Objects in O **do**
    generate distribution f_i from $\mathcal{N}(\mu_i, \sigma)$
    for each node $j \in N$ in G **do**
      draw f_{ij} from f_i
      if $f_{ij} < \Delta$ **then**
        $j_{state} \leftarrow S$
      else
        $j_{state} \leftarrow A$
      end if
    end for
    repeat
      for all j with $j_{state} = S$ AND $\Theta_j > 0$ **do**
        $\hat{f}_{ij} \leftarrow f_{ij} + \left[\frac{\Theta_j}{k_j}\right]^{(1-\gamma)}$
        if $\hat{f}_{ij} < \Delta$ **then**
          $j_{state} \leftarrow D$
        else
          $j_{state} \leftarrow A$
        end if
      end for
    until $|\{j | j_{state} = S \text{ AND } \Theta_j > 0\}| = 0$
  end for
end procedure

Master Equation We are now in the position to formulate the Master Equation for the dynamics. As already said before, two things can happen when a non-attender is connected to an attender: a) he/she becomes an attender too, or b) he/she becomes a denier who will not attend/favor the item. For these two interaction types we formally

write:

$$\begin{aligned} S + A &\xrightarrow{\lambda} 2A \\ S + A &\xrightarrow{\alpha} D + A \end{aligned} \tag{3.3}$$

Here λ denotes the probability that a susceptible node connected to an attender becomes an attender too, and α is the probability that a susceptible node attached to an attender becomes a denier. To take into account the underlying network topology of the Influence Network it is common to introduce compartments k [39]. Let N_k^A be the number of nodes in state A with k connections, N_k^S the number of nodes in state S with k connections, and N_k^D the number of nodes in state D with k connections, respectively. Furthermore we define the corresponding densities: $a_k(t) = N_k^A/N_k$, $s_k(t) = N_k^S/N_k$ and $d_k(t) = N_k^D/N_k$. N_k is the total number of nodes with k connections in the network. Since every node from N_k must be in one of the three states, $\forall t: \; a_k(t) + s_k(t) + d_k(t) = 1$. A weighted sum over all k compartments gives the total fraction of attenders at time t, $a(t) = \sum_k P(k)a_k(t)$ where $P(k)$ is the degree distribution of the network (it also holds that $a(t) = N^A(t)/N$). The time dependence of our state variables $a_k(t), d_k(t), s_k(t)$ is

$$\left.\begin{aligned} \dot{a}_k(t) &= \lambda k s_k(t)\Omega \\ \dot{d}_k(t) &= \alpha k s_k(t)\Omega \\ \dot{s}_k(t) &= -(\alpha+\lambda) k s_k(t)\Omega \end{aligned}\right\} \tag{3.4}$$

where Ω is the density of attenders in the neighborhood of susceptible node with k connections averaged over k

$$\Omega = \sum_k P(k)(k-1)a_k / \langle k \rangle \tag{3.5}$$

where $\langle k \rangle$ denotes the mean degree of the network. As outlined above, λ is the probability that a node in state S transforms to state A if it is connected to a node in state A. This happens when $\hat{f}_{ij} > \Delta$. Therefore, we have $\Delta_- < f_{ij} < \Delta$ where $\Delta_- = \Delta - (1/k)^{1-\gamma}$. From this we have $\lambda = \int_{\Delta_-}^{\Delta} f(x)dx$, where $f(x)$ is the expectation distribution. Similarly we write for $\alpha = \int_l^{\Delta_-} f(x)dx$, where l denotes the lower bound of the expectation distribution $f(x)$. A crude mean field approximation can be obtained by multiplying the right hand sides of Eq. (3.4) with $P(k)$ and summing over k, which yields a set of differential equations

$$\left.\begin{aligned} \dot{a}(t) &= \lambda \langle k \rangle s(t)a(t), \\ \dot{d}(t) &= \alpha \langle k \rangle s(t)a(t), \\ \dot{s}(t) &= -(\alpha+\lambda) \langle k \rangle s(t)a(t). \end{aligned}\right\} \tag{3.6}$$

which is later used to obtain analytical results for the attendance fraction $a(t)$.

3.3 Methods

We describe here our simulation procedures, datasets, experiments, and analytical methods.

Simulations Our simulations employ Alg. (1). As outlined in the model section, we do not change the network topology during the dynamical processes. We experiment with two different network types, Erdős-Rényi (ER), and power law (PL) which are both generated by a so-called configuration model [82]. ER and PL represent two fundamentally different classes of networks. The former is characterized by a typical degree scale (mean degree of the network), whereas the latter exhibits a fat-tailed degree distribution which is scale free. The networks are random and have no degree correlations and no particular community structure. To obtain representative results we stick to the following approach: we fix the network type, number of nodes, number of objects, and network type relevant parameters to draw an ER or PL network. We call this a configuration $\mathcal{P}$. In addition, we fix the variance σ of the anticipation distributions f_i. We perform each simulation on 50 different networks belonging to the same configuration $\mathcal{P}$ and on each network we simulate the dynamics 50 times. Then we average the obtained attendance distributions over all 2500 simulations.

Datasets To show the validity of our model we use real world recommender datasets. **MovieLens** (movielens.umn.edu), a web service from GroupLens (grouplens.org) where ratings are recorded on a five stars scale. The data set contains 1682 movies and 943 users. Only $6,5\%$ of possible votes are expressed. **Netflix** data set (netflix.com). We use the Netflix grand prize data set which contains 480189 users and 17770 movies and also uses a five stars scale. **Lastfm** data set (Lastfm.com). This data set contains social networking, tagging, and music artist listening information from users of the Last.fm online music system. There are 1892 users, 17632 artists, and 92834 user-listended artists relations in total. In addition, the data set contains 12717 bi-directional user friendship relations. These data sets are chosen because they exhibit very different attendance distributions and thus provide an excellent playground to validate our model in different settings.

Experiments **Data topologies**. We firstly investigate the simulated attendance distributions as a function of trust γ, the anticipation threshold Δ, and the network topology. For this purpose we simulate the dynamics on a toy network with 500 nodes and record the final attendance number of 300 objects. The simulation is conducted for ER and PL networks and performed as outlined in the simulations paragraph above. In Fig.(3.2) and Fig.(3.3) we investigate the skewness [119] of the attendance distributions and the maximal attendance obtained for the corresponding parameter settings. The skewness of a distribution is a measure for the asymmetry around its mean value. A positive skewness value means that there is more weight to the left from the mean, whereas a negative value indicates more weight in the right from the mean.

Fitting real data. We explore the model's ability to fit real world recommendation attendance distributions found in the described data sets. For this purpose we fix for the Netflix data set a network with 480189 nodes and perform a simulation for 17770 objects. In the MovieLens case we do the same for 943 nodes and 1682 objects and for the Lastfm data set we simulate on a network with 1892 nodes and 17632 objects. In the case of Lastfm we have the social network data as well. We validate our model on that data set by two experiments: a) we use the provided user friendship network as

simulation input and fit the attendance distribution and b) we fit the attendance distribution like in the MovieLens and Netflix case with an artificially generated network.

Mathematical analysis. We investigate the Master Equations Eq. (3.4) and Eq. (3.6). We provide a full analytical solution for Eq. (3.6) and an analytical approximation for Eq. (3.4) in the early spreading stage.

3.4 Results

Data topologies. The landscape of attendance distributions of our model is demonstrated in Fig. (3.2) and Fig. (3.3). To obtain these results, simulations were performed as described in Sec. (3.3). The item anticipation f_i was drawn from a normal distribution with mean values $\mu_i \in U(-0.1, 0.1)$ and variance $\sigma = 0.25$ fixed for all items. Both networks have 500 nodes. In the Erdős-Rényi case, we used a wiring probability $p = 0.03$ between nodes. The Power Law network was drawn with an exponent $\delta = 2.25$. The simulated attendance distributions in Fig.(3.2) and Fig.(3.3) show a wide range of different patterns for both ER and PL Influence-Networks. In particular, both network types can serve as a basis for attendance distributions with both positive and negative skewness. Therefore, the observed fat-tailed distributions are not a result of the heterogeneity of a scale free network but they are emergent properties of the dynamics produced by our model. The parameter region for highly positively-skewed distributions is the same for both network types. The parameters γ and Δ can be tuned so that all items are attended by everybody or all items are attended by nobody. While not relevant for simulating realistic attendance distributions, these extreme cases help to understand the model's flexibility.

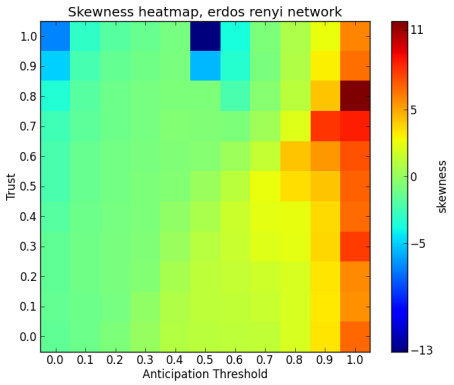

Figure 3.2: Skewness of the attendance distributions as a function of trust γ and the critical anticipation threshold Δ for Erdős-Rény networks with 500 nodes and 300 simulated items.

Fitting real data We fit real world recommender data from MovieLens, Netflix and Lastfm with results reported in Fig. (3.4), Fig. (3.5), Fig. (3.6), Fig. (3.7), and Tab. (3.1), respectively. The real and simulated distributions are compared using Kullback-Leibler (KL) divergence [64]. We report the mean, median, maximum, and minimum of the simulated and real attendance distributions. Trust γ, anticipation threshold Δ, and

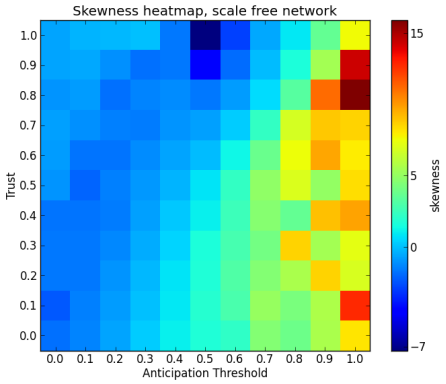

Figure 3.3: Skewness of the attendance distributions as a function of trust γ and the critical anticipation threshold Δ for power-law networks with 500 nodes and 300 simulated items.

anticipation distribution variance σ are reported in figure captions. We also compare the averaged mean degree, maximum degree, minimum degree, and clustering coefficient of the real Lastfm social network and networks obtained to fit the data. Results are reported in Tab. (3.2) and Fig. (3.8). Note that thus obtained parameter values can be useful also in real applications where, assuming that our social opinion formation model is valid, one could detect decline of the overall trust value in an online community, for example.

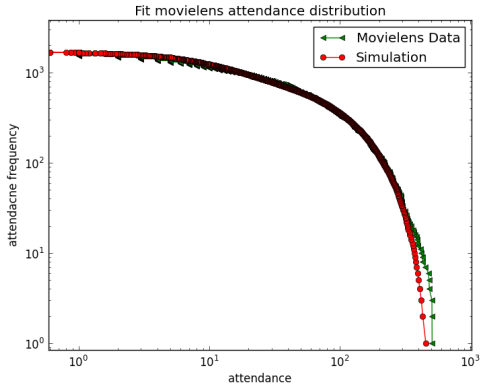

Figure 3.4: Fit of the MovieLens attendance distribution with trust $\gamma = 0.50$, critical anticipation threshold $\Delta = 0.6$, anticipation distribution variance $\sigma = 0.25$, and power law network with exponent $\delta = 2.25$, 943 nodes, and 1682 simulated objects.

Mathematical analysis. Eq. (3.6) can be solved analytically. We have $\forall t: \; a(t) + s(t) + d(t) = 1$ with the initial conditions for the first movers $a_0 = \int_\Delta^u f(x)dx$, $s(0) = 1 - a(0)$, and $d(0) = 0$. In the following we use the bra-ket notation $\langle x \rangle$ to represent

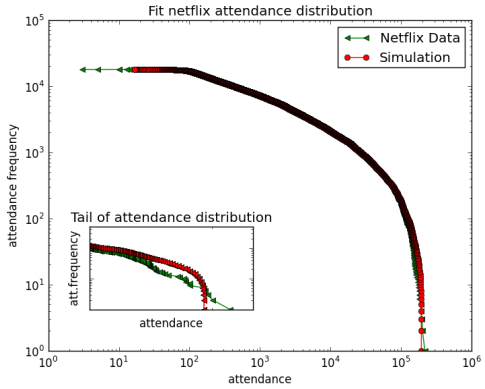

Figure 3.5: Fit of the Netflix attendance distribution with trust $\gamma = 0.52$, critical anticipation threshold $\Delta = 0.72$, anticipation distribution variance $\sigma = 0.27$, and power law network with exponent $\delta = 2.2$, 480189 nodes, and 17770 simulated objects.

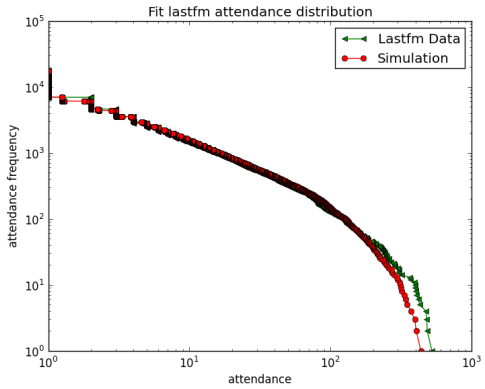

Figure 3.6: Fit of the Lastfm attendance distribution with trust $\gamma = 0.4$, critical anticipation threshold $\Delta = 0.8$, anticipation distribution variance $\sigma = 0.24$, and real Lastfm user friendship network with 1892 nodes and 17632 simulated objects.

the average of a quantity x. Standard methods can now be used to arrive at[2]

$$a(t) = \frac{(\tau \langle k \rangle)^{-1} \exp(t/\tau)}{(\alpha + \lambda)\left[\exp(t/\tau) - 1\right] + (\tau \langle k \rangle a_0)^{-1}}. \tag{3.7}$$

Here τ is the time scale of the propagation which is defined as

$$\tau = \left(a_0 \alpha \langle k \rangle + \lambda \langle k \rangle\right)^{-1}. \tag{3.8}$$

This is similar to the time scale $\tau = (\lambda \langle k \rangle)^{-1}$ in the well known SI Model [12, 83]. Eq.(3.7) can be very useful in predicting the average behavior of users in a recommender system.

Since Eq. (3.4) is not accessible to a full analytical solution, we investigate it for the early stage of the dynamics. Assuming $a(0) = a_0 \gg 0$, we can neglect the dynamics

[2]We give here only the solution for $a(t)$ because we are mainly interested in the attendance dynamics.

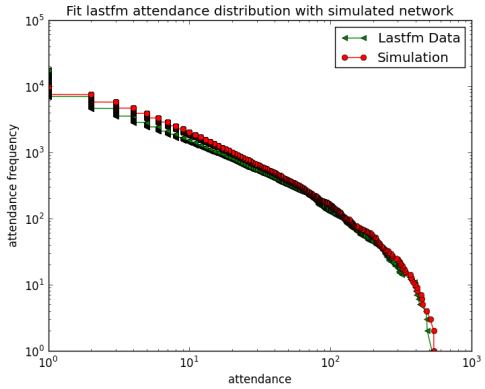

Figure 3.7: Fit of the Lastfm attendance distribution with trust $\gamma = 0.6$, critical anticipation threshold $\Delta = 0.8$, anticipation distribution variance $\sigma = 0.24$, and power law network with exponent $\delta = 2.25$, 1892 nodes and 17632 simulated objects.

D	KL	Med	Mean	Max	Min
ML	0.046	27/26	59/60	583/485	1/1
NF	0.030	561/561	5654/5837	232944/193424	3/16
LFM1	0.05	1/1	5.3/5.2	611/503	1/1
LFM2	0.028	1/1	5.3/5.8	611/547	1/1

Table 3.1: Simulation results. ML: Movielens, NF: Netflix, LFM1: Lastfm with real network, LFM2: Lastfm with simulated network, KL: Kullback-Leibler divergence, Med: Median, Mean, Max: maximal attendance (data/simulated), Min: minimal attendance (data/simulated).

of $d(t)$ to obtain

$$\dot{\Omega}(t) = \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right) \Omega(t).$$

In addition, Eq. (3.4) yields

$$\left.\begin{aligned} \dot{a}_k(t) &= \lambda k (1 - a_k(t)) \Omega(t) \\ \dot{s}_k(t) &= -(\alpha + \lambda) k (1 - a_k(t)) \Omega(t) \end{aligned}\right\} \tag{3.9}$$

Neglecting terms of order $a_k^2(t)$ and summing the solution of $a_k(t)$ over $P(k)$, we get a result for the early spreading stage

$$a(t) = a(0) \Big(1 + \tau\lambda \big(\exp(t/\tau) - 1 \big) \Big), \tag{3.10}$$

D	$\langle k \rangle$	k_{min}	k_{max}	δ	C
LFM1	13.4	1	119	2.3	0.186
LFM2	12.0	1	118	2.25	0.06

Table 3.2: Mean, minimum, maximum degree, clustering coefficient C, and estimated exponent δ of the real (LFM1) and simulated (LFM2) social network for the Lastfm data set.

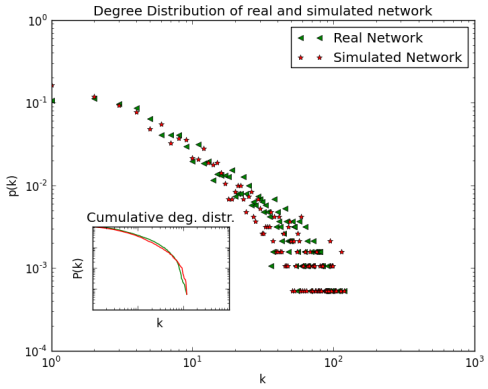

Figure 3.8: Log-log plot of real (red) and simulated (blue) social network degree distribution $P(k)$ for the Lastfm data set. Inset: plot of the cumulative degree distribution.

with the timescale $\tau = \langle k^2 \rangle / \left[\lambda(\langle k^2 \rangle - \langle k \rangle)\right]$. The obtained time scale τ valid in the early stage of the opinion spreading is clearly dominated by the network heterogeneity. This result is in line with known disease models, e.g., SI,SIR [12, 83]. We emphasize that Eq.(3.10) is valuable in predicting users' behavior of a recommender system in an early stage.

3.5 Discussion

Social influence and our peers are known to form and influence many of our opinions and, ultimately, decisions. We propose here a simple model which is based on heterogeneous agent expectations, a social network, and a formalized social influence mechanism. We analyze the model by numerical simulations and by master equation approach which is particularly suitable to describe the initial phase of the social "contagion". The proposed model is able to generate a wide range of different attendance distributions, including those observed in popular real systems (Netflix, Lastfm, and Movielens). In addition, we showed that these patterns are emergent properties of the dynamics and not imposed by topology of the underlying social network. Of particular interest is the case of Lastfm where the underlying social network is known. Calibrating the observed attendance distribution against the model then leads not only to social influence parameters but also to the degree distribution of the social network which agrees with that of the true social network.

The Kullback-Leibler distances (KL) for the simulated and real attendance distributions are below 0.05 in all cases, thus demonstrating a good fit. However, the maximum attendances could not be reproduced exactly by the model. One reason may be missing degree correlations in the simulated networks in contrast to real networks where positive degree correlations (so-called degree assortativity) are common. For the Lastfm user friendship network we observe a higher clustering coefficient $C \approx 0.18$ compared to the clustering coefficient $C \approx 0.06$ in the simulated network. To compensate for this, a higher trust parameter γ is needed to fit the real Lastfm attendance distribution with simulated networks.

We are aware that our statistics to validate the model is not complete. But we are confident, that our approach points to a fruitful research direction to understand recommender systems' data as a social driven process.

The proposed model can be a first step towards a data generator to simulate bipartite user-object data with real-world data properties. This could be used to test and validate new recommender algorithms and methods. Future research directions may expand the proposed model to generate ratings within a predefined scale. Moreover, it could be very interesting to investigate the model in the scope of social imitation [77].

Chapter 4

Emerging cohesion and individualization in collective action: a co-evolutive approach

4.1 Introduction

Populations under stress can present surprising and powerful behaviours. From the uprising of spontaneous social protest against authoritarian regimes leading to their fall, to the emergence of unexpected solidarity among victims of terrorist attacks (see September 11th 2001 or March 11 2004 Madrid, for instance), we could mention a long list of situations where conflicting scenarios made social cohesion in human groups to emerge.

Some scholars studying different empirical cases have stressed the importance of pre-existing informal social networks in such an emergence. The role of personal networks as mobilization contexts in East-German and October's Serbian revolutions is the main interest of Opp and Gern in [86], and Araya in [7], respectively. Opp and Gern present a complete analysis on the incentives of individuals to join Leipzig's Monday demonstrations in 1989. They uncover the little influence of opposition organizations in comparison with that of personal networks of friends. Araya introduces the concept of *cooperative cascade* to describe the link between ego-centric networks' characteristics and macroscopic mobilization phenomena. Kinship structures have also been considered. In [81], Murphy studies their relationship with warfare organization patterns of a Brazilian Indian group, the Mundurucú. The Mundurucú were settled in several apart villages, spread along the upper Tapajós River. However, the setup of war parties evidenced a strong relationship among these communities, otherwise unobservable. Murphy concludes that intercommunity cooperation in warfare was facilitated by cross-cutting ties of residential affinity and affiliation by descent.

Other authors have observed that there exists also an influence in the opposite direction, that is, of mobilization over informal social networks. In [44], Gould analyzes the insurgent activity during the Paris Commune in 1871, which sprung after a mixture of political, economic and war crises. In this paper, as in later works [45, 46], he settles that organizational networks and pre-existing informal networks interacted in the mobilization process. As Gould points out, mobilization does not just depend

on existing social ties; it also creates them. Although members of a protest organization may have joined because of a pre-existing social tie to an activist, they eventually also formed new social relations while participating in collective protest. In other words, opinion affinity manifested by joint mobilization led to the formation of new ties among individuals.

Summarizing, in order to analyze the emergence of cohesiveness in conflictive scenarios, we need to observe the dynamic interplay between structural and cognitive components of social cohesion during the period of activity. Notice that this conclusion illustrates perfectly what Giddens defined as "duality of structure". According to him, the social structure is simultaneously the product and the constraining environment of social action and, therefore, these two entities cannot be studied separately [38].

This paper aims at addressing this interplay in a quantitative way by modeling the co-evolution between individual behaviours (opinions) and social networks [66]. Different quantitative approaches have been developed to study the evolution of the social structure under the influence of local dynamics (e.g. strategic network formation, network evolution models and exponential random graph models) [54, 106]. However, it is only very recently that we have started to see simulation-based studies addressing the co-evolution of structure and dynamics [47, 92]. Specifically, our model's dynamics, based on the proposal by Holme and Newman in [53], make the system to evolve in a twofold way: individuals become likeminded because they are connected via the network (change is induced by the structure) and they form network connections because they are like-minded (structure undergoes change).

The model reveals to be a good framework to reproduce and study the evolution of social cohesion in a population submitted to sudden changes on its environmental conflict level. Moreover, a detailed analysis of its dynamics uncovers the counterintuitive effect of noise and the importance of the social structure at the microscopic and mesoscopic structural level.

Social movement behaves in a regular pattern; from the institutional (macro) level to the individual (micro) sphere through the (meso) level of networks and vice versa [28]. This interaction at the meso level is complex and it is constituted both by processes of selection on the part of individual and influence by groups [103]. In other words, a mobilization begins with a mobilization potential which depends both on macrostructural factors such as demographic, economic or ideological variables and individuals predispositions and social networks structures in which they are embedded, who, in turn, change their connectivity thus affecting social groups' structure and the macrostructural framework. In particular, our analysis reveals that a moderate rate of noise (here seen as an individualistic trait) can enhance the social cohesion of a population by enabling cross interactions among the groups forming it.

The remainder of the paper is organized in three sections. The second section is devoted to the detailed description of the model, making an special insight on the influence of the social noise. Simulation results are presented and discussed in section 3, focusing specially on the role of the different topological levels. Finally, the last section summarizes the work and proposes further extensions.

4.2 Cohesion analysis through a coevolutive model

4.2.1 The model

We consider a population of N agents, connected through a variable number of undirected (bidirectional) links. Each agent i presents a h_i value, corresponding to his location in a continuous lineal social space of size L (proportional to N). Here, h_i could be seen as an opinion or positioning of individual i in relation to a certain topic (related to religion or politics, for instance). Notice that this approach has been commonly adopted in the well established literature about continuous opinion modeling [1, 29, 96].

Initially the h values of all agents are assigned randomly, following a uniform distribution along the lineal social space. Besides, the initial arrangement of the edges correspond to a topology with the same structural properties than real social networks (like large clustering coefficient and positive degree correlations, for instance). To construct such a scenario, we use a class of models proposed in [20], which are able to grow up networks with social-like macroscopical (global) properties from a microscopical (individual) definition of the linkage probability between two agents. The key element of that definition, is the social distance between the two individuals in a social space of a certain dimension $d_{\mathcal{H}} \geq 1$. By social distance, here we mean *"the degree of closeness or acceptance that an individual or group feels towards another individual or group"*[20]. Since our social space is lineal, here we use a simplified expression of the linkage probability with $d_{\mathcal{H}} = 1$

$$r(h_i, h_j) = \frac{1}{1 + \left[b^{-1}|h_i - h_j|\right]^{\alpha}} \tag{4.1}$$

where $|h_i - h_j|$ corresponds to the social distance, b a parameter controlling the length scale of the lineal social space, and α quantifies the homophily, which is everyone's preference to establish and maintain relations with people that have any common characteristic with (cultural background or political feelings, for instance) [75]. Therefore, given a certain social distance between two agents, different combinations of b and α values lead to different link probabilities, in such a way that the higher the b and the lower the homophily, the larger the probability of connection.

Our model evolves from the initial scenario in a twofold way, by redefining both the topology of the network and the positioning of the population of agents in the social space. Based on a co-evolution model proposed by Holme and Newman in [53], the two main mechanisms driving this co-evolution process are the rewiring of links and the imitation of h values among agents. Additionally, we have incorporated a third mechanism that reproduces slight shifts on each one's opinion or social position, induced by individual circumstances and daily life experiences, which usually modify individuals' knowledge in a subtle but continuous way. This third mechanism is necessarily external, since these particular characteristics are different for each individual, and don't depend on any other parameter of the model. Notice that, at the mid-long time range, these slight but continuous shifts can change significantly the social distance among two individuals, separating two agents that were once very close in the lineal social space or, on the contrary, approximating them enough to favor the cre-

ation of a new link. Consequently, the accumulation of these microscopical changes can modify the whole macroscopical scenario, by disrupting both the distribution of agents' positions along the social space and their connectivity. Taking into account this disrupting effect, and in alignment with previous literature introducing noise in a similar way [71], we have denoted this third mechanism of the dynamics as *individualization noise*.

These three mechanisms (rewiring, imitation and individualization noise) are integrated within the co-evolutive dynamics of the model, consisting on the repetition of the following two steps:

1. Select an agent x at random and decide, with equal probability, whether to apply rewiring or imitation.
   - The rewiring consists on a redefinition of all links of node x using the expression in (1).
   - Imitation is implemented by selecting randomly a neighbor y of node x, and setting h_y equal to h_x.
2. Introduce the *individualization noise* by summing up a random quantity **to the h value of each agent in the population. This random quantity is obtained from a Gaussian distribution (with mean=0.0 and variance=1.0) multiplied by a noise *magnitude* or *strength* factor n.**

Fig. 4.1 illustrates these dynamics. At each time step, the system evolves following one of the two possible branches of the diagram (imitation plus *individualization noise*, or rewiring plus *individualization noise*) with the same probability. Notice that this probability (or, in other words, the relative proportion at which imitation and rewiring occur) has been found to play an important role in this kind of co-evolutionary models. Vazquez and co-authors, for instance, showed that there is a phase transition towards fragmentation varying this probability [108]. In order to discard potential effects of such a transition in our particular case, we have performed additional simulations (not shown). **These simulations show that the rewiring probability used in this paper ($p = 0.5$) is such that it is well below the critical value for fragmentation.**

After a certain number of time steps, the system reaches a *steady-state*. In our context, this means that both the topology and the distribution of individuals' social positions along the space remain stable. The concrete topology and distribution of social positions reached at each possible steady-state depend, as we will show in the next subsection, on the strength of the *individualization noise*.

4.2.2 Effect of *individualization noise*

An important issue to deep in at this point, is the influence of the *individualization noise* over the evolution of the model. Different kinds of noise have been reported to influence enormously different opinion and cultural dynamics models.

One of the most relevant examples in the literature is [62], where the authors show that the stable cultural diversity achieved by Axelrod's model of dissemination of culture [9], can be easily fused into cultural homogeneity by introducing small random

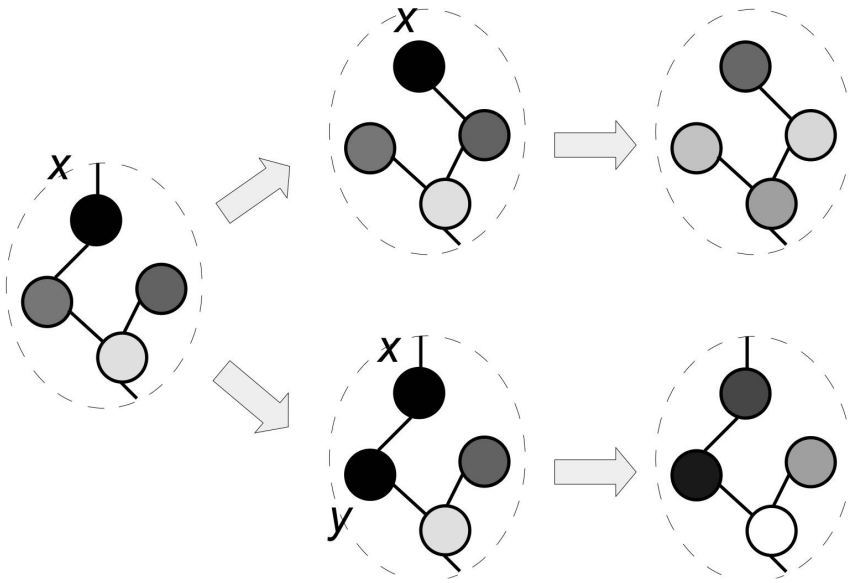

Figure 4.1: An illustration of our co-evolutive dynamics, where colors indicate the h value of each individual. At each time step, the system evolves following one of the two branches. The upper branch correspond to a rewiring (of $x's$ links) plus a shift of all positions, and the lower one to imitation (y imitates x) plus position shifts.

variations in agents' cultural features (what they call *cultural drift*). This result is explained by the fact that cultural drift can eventually make an agent to share cultural traits with agents belonging to completely different groups, allowing crossed social influence among previously isolated agents and, therefore, leading towards cultural homogeneity. After this observation was made, robustness to noise has become an important requirement for models trying to reproduce the emergence of opinion diversity [26, 31, 32].

More recently some authors have started to introduce noise in models to account for opinion individualization. Pineda and coauthors proposed a new version of Deffuant's bounded-confidence model where noise is used to model individuals' *free will* [93]. Specifically, in that model agents' were given the opportunity (with a certain probability) of changing their opinions to a randomly selected position in the whole opinion space. The authors found that the noise defined in such a way was able to induce a transition between a disordered state (where opinions were distributed uniformly) and an ordered one (where opinion clusters emerged). Similarly, Mäs and *et al.* used a noisy model to show how moderate rates of individualization can lead to opinion clustering [71]. In this case the noise plays the role of a uniqueness-seeking effect, which counteracts the general tendency of agents' opinions towards consensus around the average opinion in the population. More concretely, the noise is defined by a normally distributed random variable which standard deviation is higher the more homogeneous is the social context of the opinion holder (i.e. the more similar to her are the opinions of the other agents in the population). Finally, this same noise-based mechanism is used in a subsequent paper to explain the persistence of social differentiation in groups and organizations [70].

Noise definition in our model is somehow related to the two described above. On one side, it is independent of the social context of the opinion holder (as in [93]). On the other hand, in accordance with [71], it is defined by a normal distribution since small opinion changes are much more likely than large ones. Moreover, also as in [71], we keep all agents' positions within the interval [0,L] by not applying individualization noises if such boundaries would be crossed otherwise.

Being our individualization noise defined by a Gaussian distribution with a fixed mean and variance, we center our attention on the unique parameter that can be tuned: its magnitude. In the context of our study, the magnitude corresponds to the average range of changes experimented by individuals' social position due to the *individualization noise*. Strong *individualization noise* implies sudden changes of individual's social positions along the social space. On the contrary, weak noises correspond to quite stable opinions.

Taking this into account, we can easily predict the behavior of the model for extremal values of the noise magnitude. On one side, too much *individualization noise* would result in a noise-dominated scenario, where agents would be almost completely isolated due to the difficulty to maintain links among them. On the other hand, too low noise intensity would exercise no significant effect over the dynamics, which would be controlled by the other two mechanisms (imitation and rewiring). Keeping this in mind, some questions arise: what are we to understand as "too weak" or "too strong" noise? And, how does the noise influence the dynamics for intermediate strength values between these limits?

In order to address these questions, we have analyzed the influence of different noise magnitudes over three topological measures, namely the density within clusters (ρ), the average degree ($< k >$) and the average Clustering Coefficient (Cc). In Figure 4.2, we present the evolution of these measures for a given set of initial conditions and different values of the noise strength. For extremal values of the noise magnitude, results corroborate predicted behaviors. Unexpectedly, however, we observe that the case corresponding to an intermediate noise strength leads to steady-states with the highest average degree.

Such a surprising result can be related, in our particular case, to the capacity of a moderate individualization noise to introduce heterogeneity within the different groups. This internal diversity favors the inter-group linkage without breaking them into isolated agents.

Let's explain this argument more accurately. When the individualization noise is weak or absent, the combined action of imitation and rewiring leads the model to a steady-state where individuals tend to coincide in a unique h value (social position) and, therefore (because of the rewiring action), to conform a unique connected component. On the contrary, when the magnitude of the individualization noise is extremely high, differences between h values of agents (social distances among them) grow such quickly that cannot be counteracted by the imitation mechanism and, when those distances are too large to maintain links between neighbours, groups are progressively dissolved towards a completely disconnected scenario. In an intermediate situation, the noise intensity is high enough to maintain a wide variety of h values, but the differences introduced among these values are small enough to keep agents linked and, in some cases, to establish new links with agents belonging to other groups.

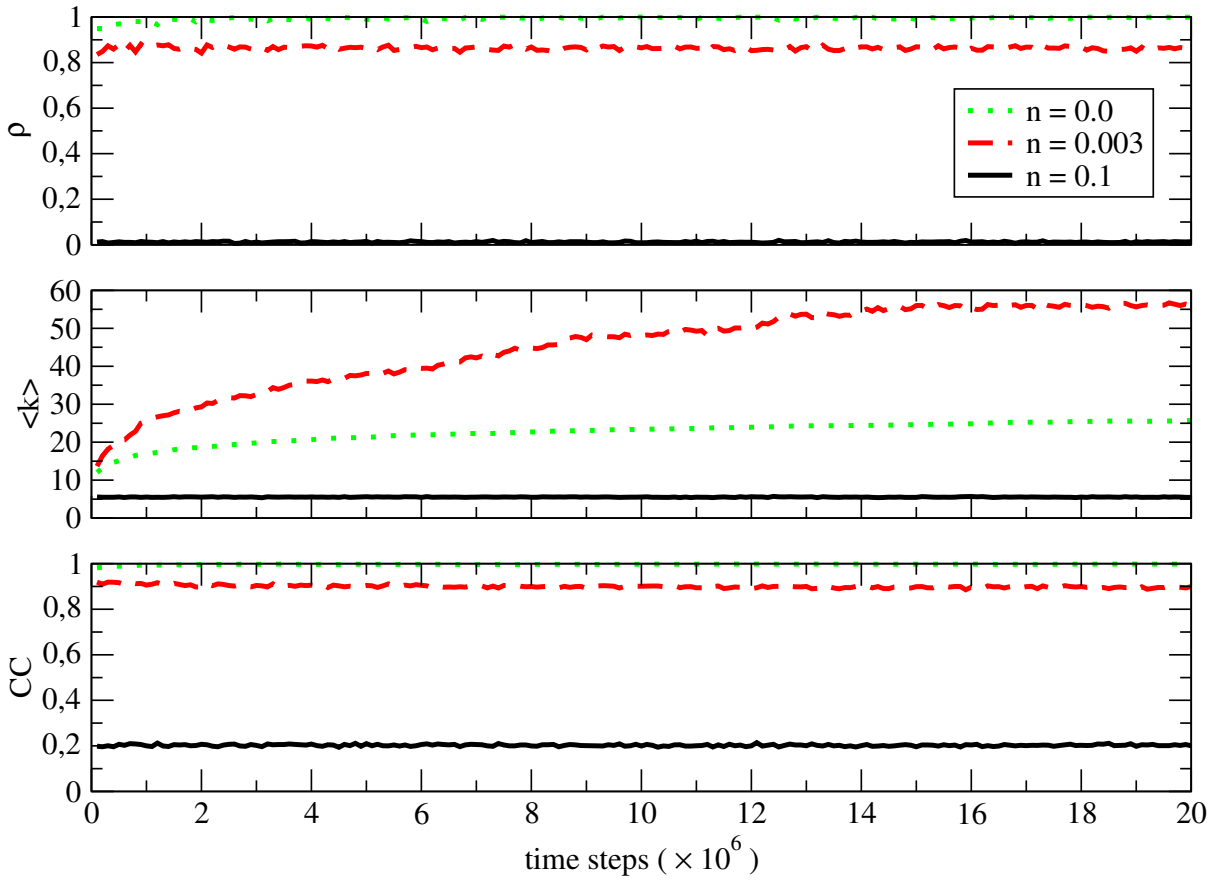

Figure 4.2: Influence of noise magnitude over model dynamics. Evolution of the density within clusters (ρ), the average degree ($< k >$) and the average Clustering Coefficient (Cc), for **two** different values of the noise strength (representative of strong and intermediate noise strength). The case without noise **(**$n = 00.0$**)** is also shown, for comparative purposes. Population size N, size of the opinion space L and homophily α were set to 1000, $N/5$ and 6, respectively. Results were averaged among 25 independent realizations.

Note that a moderate individualization noise in our model and Klemm and coworker's cultural drift in [62] act in a parallel way, since both of them facilitate the interaction between otherwise isolated components. Sort to say, they 'liquefy' an stable scenario composed by separated components (cultural regions for the cultural drift, opinion groups in our case).

4.3 Results and discussion

4.3.1 Experiment

By using the described model as a framework, we have conducted a simulation experiment to study social cohesion and its interplay with extremal changes on the social environment. Such an experiment comprises two crisis cycles (sudden increases of the social temperature followed by longer reactionary periods). Each one of the crisis cycles has consisted on a short period (about 50000 time steps) of high social temperature, followed by a fall to extremely low temperature (reproducing an habitual reactive behavior of populations after an emergency situation) and, finally, a progressive recovery towards normality.

In order to run the described experiment, we need to be able to simulate different social temperatures in our model. Such changes on the social temperature, has been modeled as variations on the value of the b parameter (the one controlling the length scale of the social space). This solution can be justified as follows. When some kind of emergency strikes a population, social distances that separate individuals do not change, but the necessity to face the new critical scenario makes them less important than in a quiet situation. This temporal relativization of social distances is nothing but a change on the length of the scale they are 'measured' against. Consequently, an appropriate way to introduce in our model the effect of emergencies and posterior relaxations of the conflict level, is to increase the value of b (making distances relatively smaller) and, after a relatively short number of time steps, decrease it back. In our case, the b values chosen to represent each period are 0.5 for 'normal' social temperature, 2.0 for highly conflictive situations and 0.25 - 0.35 for the reactionary intervals.

Furthermore, we also want to monitor the evolution of social cohesiveness under these environmental changes. For this purpose, we have used three different macroscopical observables, namely: the average degree $\langle k \rangle$ (average number of neighbors), the clustering coefficient (a weighted measurement of the number of triangles) and the number of disconnected components or independent groups G composing the whole network. While first and second parameters signal intra-group cohesion, the third one corresponds to inter-group cohesiveness. Note that, taken jointly, these three are good indicators of the social cohesiveness, since the more cohesive is a population, the higher are their average degree and clustering coefficients, and fewer separate groups it presents.

When looking at the behavior of these observables during the experiment, shown in Fig. 4.3, we observe two phenomena. First we notice that, for the same value of b, the social cohesiveness after each emergency situation is higher than before them. Second, we observe a memory effect on the cohesion of the system in the period between

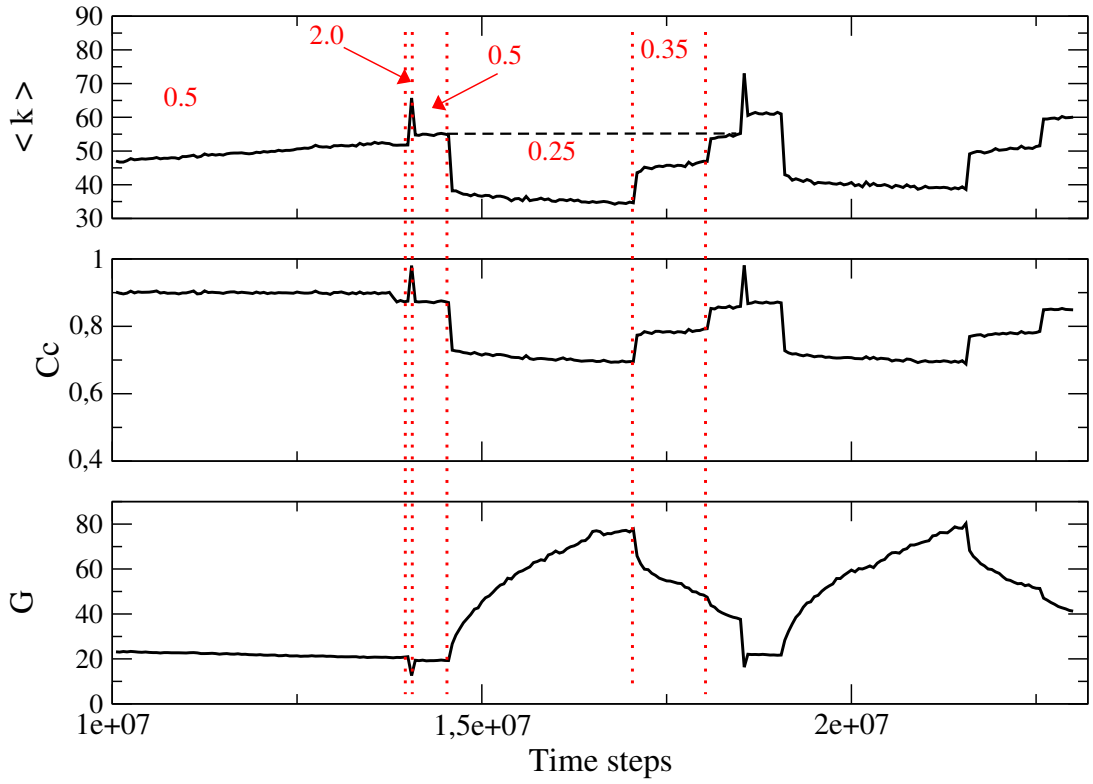

Figure 4.3: Evolution of social cohesiveness during the experiment. Vertical dashed lines in red indicate regions delimited by their b value, which are indicated also in red. Values of sizes N and L, as well as homophily α, were kept as in Fig. 2. The noise magnitude was set to the intermediate value 0.003. Two important phenomena are observed: An increase on the average cohesion after each crisis, and a memory effect in the period between crises (represented here by a horizontal dashed line in black). Results were obtained by averaging 25 independent realizations.

crises. Although the cohesiveness diminishes as a response to social temperature cooling, when the situation comes back to normality, the cohesiveness also recovers its "normal" value (that one corresponding to $b = 0.5$ just after the crisis).

The first result agrees with the observation made in the introduction in the sense that the structure of the system changes during the conflict period. Moreover, it can be positively contrasted with observations of other real social systems. When a population has been submitted to a stressing situation, it is quite usual to find higher levels of cohesion than before the crisis. In some sense, this phenomenon could be seen as a sort of *reminiscence* of the high rates of cohesion characterizing the emergency situation.

4.3.2 Analyzing mesoscopic and microscopic dynamical aspects of social cohesion

Up to this point, our model has revealed its capacity to reproduce how changes on the social temperature (which is a macroscopical variable related to the social environment) induces changes on the cohesiveness of a population of individuals (here

measured in terms of macroscopical observables).

Nevertheless, in the introduction we have pointed out that the analysis of the concept of social cohesion from a dynamical viewpoint demands a more complete scope of the problem, including also the behavior of different variables at meso and micro levels during the conflict period. In order to deep in this issue, we have studied how our model's dynamics modifies the distribution of agents' positions (h_i values) along the lineal social space and, consequently, how the social structure of the population is transformed.

In general, when plotting the distribution of agents' opinions in the social space at a steady-state (see Fig. 4.4 for two particular examples), we find that agents are grouped around certain positions of the space, and that there are quite regular separations among these concentrations. Taking into account the dependence of the linkage probability on the social distance, we deduce that these concentrations of opinions in the social space correspond, structurally speaking, to groups of agents densely connected. Besides, the observed separations tend to a unique value that we have called *critical social distance* $d_c(h_i, h_j)$, which is the maximum social distance at which connectivity between two groups of agents is possible. In other words, is the distance making the link probability close enough to 0 as to have just one expected link between the two groups. **Notice that, in accordance with this definition, d_c (and the corresponding r) depends very much on the particular scenario. For instance, if we had one single agent on one side and the rest of the population ($N-1$) on the other, r would take the value $1/(N-1)$. However, for a scenario with two groups of equal size $N/2$, we would need $r = 1/(N^2/4) = 4/N^2$.**

We propose the following formalization for the *critical social distance*:

$$d_c(h_i, h_j) = \lim_{r\to 0} d(h_i, h_j) = \lim_{r\to 0} b \sqrt[\alpha]{\frac{1}{r(h_i, h_j)} - 1} \approx \frac{b}{\sqrt[\alpha]{r(h_i, h_j)}} \tag{4.2}$$

From this definition, it is straightforward that links are established only among agents separated by a distance smaller than $d_c(h_i, h_j)$. Moreover, the combined effect of imitation and rewiring makes that any agent located in a social position shorter than $d_c(h_i, h_j)$ from any group tend to link to that group, and that two groups tend to merge if they are near enough from each other. Consequently, in the steady-state not only the distance among groups, but also their number and size, is related to the critical social distance. The larger the $d_c(h_i, h_j)$, the fewer separated groups and the larger the distance among them.

Furthermore, by taking a look to expression (4.2), we realize that *the critical social distance* depends on b. Since this variable controls the *social temperature* in our model, we can trace a causal path from variations on the social temperature to structural and knowledge changes experimented by the population during a crisis period. With this idea in mind, we can interpret the behavior of the cohesiveness during the experiment (shown in fig 4.3) in terms of reductions and increases of the critical distance, induced by changes on b (that is, the social temperature).

At the beginning of the experiment, before the first crisis, the critical distance is defined by the original b value (0.5). When the b value becomes 2.0, the critical distance

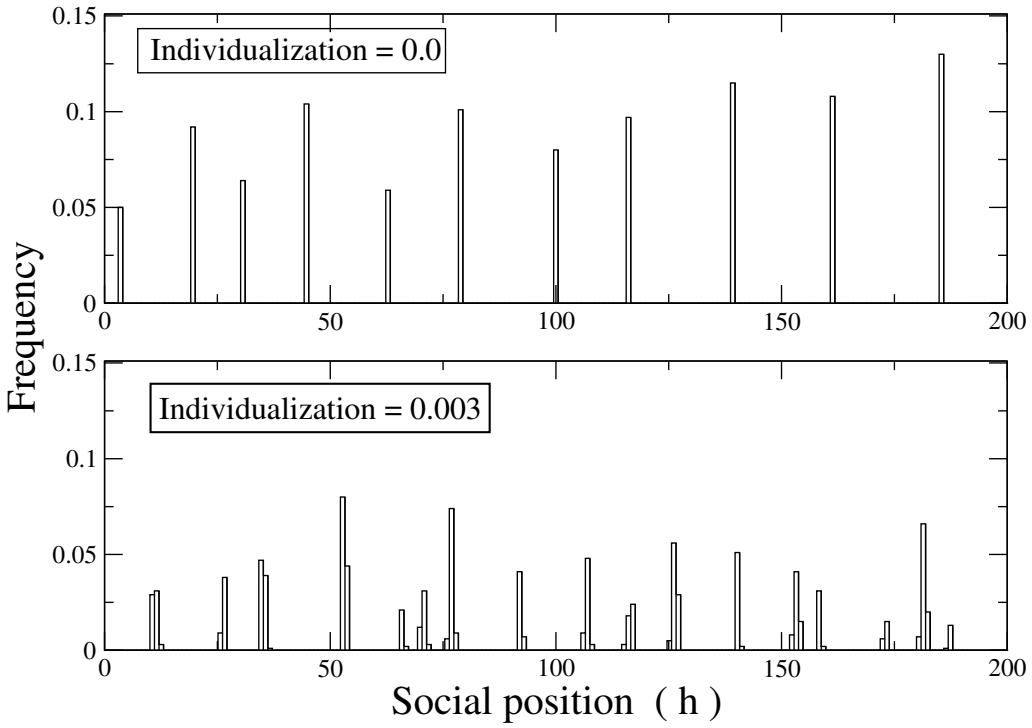

Figure 4.4: Distribution of agents' positions along the lineal social space, in a steady-state, without individualization noise (top) and with a individualization noise of magnitude 0.003. Although both cases present quite regular separations between groups (see text for an explanation), the internal distribution of each group differs. In the bottom case, we appreciate the heterogeneity within groups introduced by the individualization noise.

also increases and, consequently, all agents come across other ones that were previously out of their range. Globally, this means that the population tend to reorganize into fewer but larger groups, whose opinions are separated each other by greater social distances. However, as this process is interrupted abruptly (due to the briefness of emergency situations), some agents are 'surprised' halfway between various groups. After a short transitory period, a new steady-state is reached. In this new stable scenario, agents conserve many neighbors of the period before the crisis and have incorporated new ones due to those agents bridging different groups after the emergency. Consequently, the resulting groups are larger than before the crisis and, because of their high internal connectivity, the average degree and the clustering coefficient also keep higher. This phenomenon is what we have previously called *reminiscence* of the crisis over the social cohesiveness.

At this point, agents are 'trapped' within their groups (i.e., their opinions are too much different from those of other groups to establish cross-links). Moreover, in this second stable period, the individualization noise plays a central role by opening very little internal discrepancies between members of the same group, that allow the creation of new groups when the social temperature gets 'colder' (b drops down to 0.25). Later, as the population recovers its 'normal' social temperature (and, therefore, the b value increases again), the critical distance grows up and little groups tend to merge and recover the stable configuration reached just after the crisis, presenting the second phenomenon pointed above, a memory effect. Finally, during the second cycle, the system presents the same behavior than in the first one: A higher cohesiveness than before and a memory effect.

4.4 Conclusions

In this work we have developed a simple model as an analytic tool to explore a dynamic perspective of the concept of *social cohesion*, integrating the already-stated structural component [80] with a cognitive, cultural one. Given a certain initial scenario, the model evolves under the influence of the conflict level of the environment by redefining, simultaneously, the social structure and the knowledge or opinions (represented as positions in a social space) of a population of agents. We argue that, beyond static perspectives, the social cohesion of a population should be expressed in terms of these changes experimented both at structural and cognitive dimensions as a response to conflict increases.

By means of only three simple mechanisms, the dynamics of the model reproduces the behavior of real social populations under a highly conflictive situation. We have showed this in a twofold way. First we have studied numerically how changes on a variable of the system representing the *social temperature* (degree of conflict) conditions the evolution of three observables than can be easily related to social cohesion (average degree, clustering coefficient and number of isolated components). Second, we have deepened in dynamic aspects of social cohesion by tracing the causal path among different topological levels. Changes on social temperature happen at an institutional level, influencing relationships among agents (microscopical level), and these changes at the individual level modify the size and composition of groups conforming the

social population (mesoscopic or intermediate level).

Although having demonstrated its utility as a tool to analyze the concept of social cohesion, there are some aspects of the model that could be explored in order to make it closer to particular case studies. In the following, we point out two of these possible extensions of the model.

The initial conditions of our experiment, determined by a topology and a distribution of agents' opinions, can be defined in many different ways. In this case, we have chosen a simplistic initial scenario (synthetic social-like topologies and a uniform distribution of opinions) in order to show that, even starting with such simple conditions, our model is able to reproduce certain phenomena related to social cohesion and its dependence on variations on the social temperature. Nevertheless, each one of the two components of the initial scenario can be modified separately. For example, we could use an empirically obtained social network as the initial topology, but we could also start out the experiment with a distribution of opinions representing a scenario of preexistent coalitions or opinion groups.

Another possible extension of the model is related to the observables used to quantify the evolution of the social cohesion. Although the three structural observables used in this work are too much simple to represent population's cohesion separately, analyzing the evolution of their behaviors jointly has helped us to understand the dynamical processes taking place in the model. Nevertheless, for the sake of simpleness and clarity, it would be interesting to define a unique (necessarily more complex) structural observable, based on previous studies like [113] and [80]. Furthermore, in accordance with the aim of this work of enriching the structural approach to social cohesion with a cultural component, it would also be interesting to define an observable related to the distribution of opinions in the social space (based on the largest social distance in the system, for instance).

Finally, current on-line communication platforms open new sociological research possibilities. The eruption of social networking sites like Twitter or Facebook and their undiscussed key role in recent civil mobilizations (wave of protests in the Arab world, the M15 movement in Spain [21, 42]) and riots (England, summer 2011) provide an unprecedented chance to empirically test theoretical models relying both on underlying bond topologies (who holds stable relations with whom) and on information dynamics (who is actually communicating with whom).

Chapter 5

Overlay management for fully distributed user-based collaborative filtering

5.1 Introduction

Offering useful recommendations to users of fully distributed systems is clearly a desirable function in many application domains. Some examples for larger efforts towards this goal are the Tribler platform [35] and more recently the Gossple project [59]. A fully distributed approach is also more preferable relative to centralized solutions, due to the increasing concerns over privacy.

However, the problem is also extremely challenging. Apart from the fact that centralized recommender systems—although working reasonably sometimes—are still far from perfect, offering good recommendations in fully distributed systems involves a number of special problems like efficiency, security and reliability, to name just a few.

In this work we focus on a class of recommender systems, the so called user-based collaborative filtering algorithms that are fairly simple, yet provide a reasonable performance [3]. The key concept is a similarity metric over the users, and recommendations are made on the basis of information about similar users.

This idea also naturally lends itself to a distributed implementation, as it can be easily supported by similarity-based overlay networks as a simple service, that also have applications in other domains such as search. Indeed, many distributed protocols from related work follow this path in some way or another.

In this work we would like to shed light on the effects of the basic design choices in this domain with respect to recommendation performance, convergence time, and the balancing of the network load that the system generates during its operation.

Our contribution is threefold. First, we draw attention to the potential load balancing problem in distributed systems that manage similarity-based overlays for any purpose including recommendation or search. Second, we propose novel algorithms for similarity-based overlay construction. Third, we perform extensive simulation experiments on large benchmark datasets and compare our set of algorithms with each other and with a number of baselines. We measure prediction performance, examine its convergence and dynamics, and we measure load balancing as well.

The outline of the chapter is as follows. In Section 5.2 we discuss related work, in Section 5.3 we analyze three widely used benchmark datasets (to be used in our experiments) from the point of view of potential problems in a distributed setting. Section 5.4 discusses the set of algorithms that we test empirically in Section 5.6 assuming the system model described in Section 5.5. Section 5.7 concludes the chapter.

5.2 Related Work

First we overview relevant ideas in recommender systems in general, and subsequently we discuss related work in the fully distributed implementations of these ideas, as well as additional related work that are based on similar abstractions.

A recommender system can be viewed as a service which supports e-commerce activities by providing items of interest for the users [94]. These algorithms are often centralized and Web-based operating on huge amounts of data—mainly on the previous ratings of the users. The algorithms which are based on the previous ratings of other *similar* users follow the so-called collaborative filtering (CF) approach. They are based on the simple heuristic that people who agreed (or disagreed) in the past will probably agree (or disagree) again. Thus, the predicted rate of an unseen item for a given user can be estimated on the basis of the rates of other users with *similar tastes*.

In the field of CF algorithms there exist numerous approaches. *User-based* approaches try to model the rating of a given item for a user by an aggregation of ratings of other users on the same item [3]. Although these approaches are very simple and intuitive, they provide a relatively good performance [52]. User-based CF algorithms are modular, hence they can be used with different aggregation methods and similarity metrics. One widely-used aggregation method is

$$\hat{r}_{u,i} = \frac{\sum_{v \in N_u} s_{u,v} \left(r_{v,i} - \bar{r}_v \right)}{\sum_{v \in N_u} |s_{u,v}|} + \bar{r}_u \tag{5.1}$$

defined in [98], where $r_{u,i}$ and $\hat{r}_{u,i}$ denote the known and the predicted rate of item i by user u, $\bar{r}_u$ and N_u denote the average rate and the neighbor set of user u, and $s_{u,v}$ measures the similarity between user u and v (e.g. Cosine similarity [3] or Pearson similarity [3] can be employed).

Our preliminary experiments showed that (among several variants) the aggregation method in (5.1) combined with the Cosine user similarity gives the best performance on our particular benchmarks. Since the focus of the present work is not recommendation performance per se, but the analysis of several distributed implementations of the basic idea of user-based CF, we fixed these methods in our experiments.

We should mention that there are numerous other approaches for recommendation such as the ones based on machine learning [16, 90], matrix factorization [105], generative models [65], clustering [84, 90], and dimension-reduction [16, 41].

Moving on to distributed methods, we emphasize that we focus on P2P recommendation, and not on parallel implementations of centralized recommender techniques (such as matrix factorization, etc.). We consider only works that go beyond a simple idea and present at least some evaluations on benchmarks.

The largest group of methods define an overlay network based on some sort of similarity, and define a recommender algorithm on this network. For example, [94] and [25] follow this approach, although the overlay construction itself is not discussed or it is assumed to be done offline. The recommender algorithms then perform a search in this overlay up to a certain depth or up to a certain level of similarity, and aggregate the matching users with a standard method.

A slightly weaker approach is described in [107], where only a random network is assumed and the recommendation problem is treated as a search problem where a node needs to find similar users using a flooding based unstructured search.

A somewhat surprising result is described by Bakker at al [10], where they argue that in fact it is enough to take a random sample of the network and use the closest elements of that sample to make recommendations. Our results are consistent with this observation, although we describe better and equally cheap alternatives.

A more sophisticated approach is described by Bickson et al [14]. They define recommendation as a smoothing operation over a social network, which is expressed as a minimization problem using an objective function that expresses the requirements for the recommendation. The problem is solved by using an iterative method. Unfortunately no results are given on recommender system benchmarks due to the slightly different formulation of the basic problem.

It is of course possible to apply distributed hash tables [49]. Here, users are stored in a hash table and they are indexed by (item, rate) pairs as keys. Using this data structure, the users for a given item and rate are available from the distributed hash table (DHT) on demand. This method is not scalable if there are many recommendations to be made in the system, since the necessary information is not always available locally.

One of the most detailed studies on distributed recommender systems with performance evaluation can be found in [111]. The proposed models were implemented on the basis of the BUDDYCAST [95] overlay management service, which is the main overlay management method of the Tribler file sharing protocol [35]. We used our own implementation of this model as a baseline method, since the original study [111] did not carry out load balancing measurements.

Finally, although not directly related to the recommender systems, the area of exploiting semantic proximity for search also involves building overlay networks based on node similarity and therefore our algorithms and observations are relevant in this area as well. Examples of research in this area are described in [4, 36, 59, 109].

5.3 Interesting Properties of CF Datasets

In our simulations we applied three different benchmark datasets, namely the MovieLens [52] dataset, the Jester [41] dataset and the BookCrossing [118] dataset. In this section we introduce these benchmarks and show some of their properties that raise interesting—and so far largely overlooked—problems in distributed environments.

Table 5.1 summarizes some basic statistics of our datasets. In the case of MovieLens we used the official r_a partition so that its evaluation set contained 10 ratings per user. For Jester and BookCrossing we produced the evaluation set as proposed in [10]: we withheld 6 ratings from the training set where possible (if the user under consideration

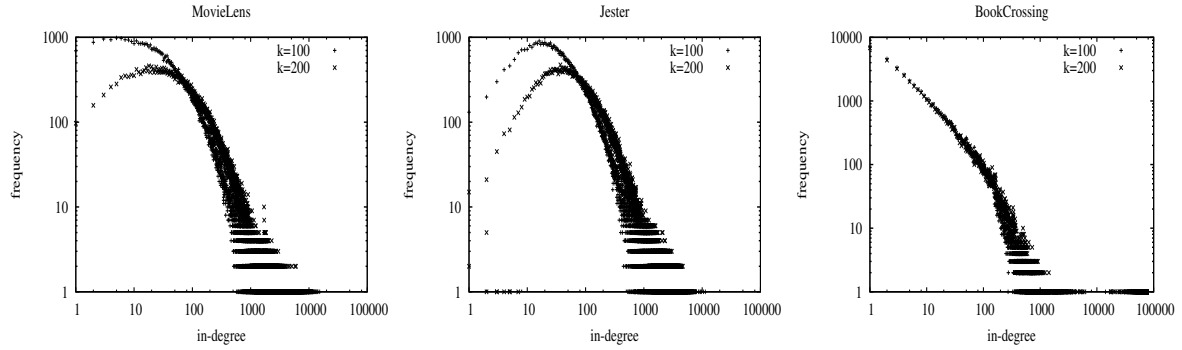

Figure 5.1: In-Degree Distribution of Benchmark Datasets

had at least 6 rated items). In this table '# items $\geq$' means the minimal number of items rated by some user. Sparsity denotes the ratio of existing and possible rates in the training sets. The value MAE(med) is a trivial baseline for prediction performance; it is defined as the mean absolute error (MAE) computed on the evaluation set using the median-rate of training set as a prediction value. Clearly, a very significant difference can be found in properties related to sparsity. This will have significant implications on the performance of our algorithms, as we show later.

As mentioned before, in distributed settings one suitable and popular approach is to build and manage an overlay that connects similar users. This overlay can be viewed as a graph where each node corresponds to a user and there is a directed edge between user A and B if and only if user B belongs to the most similar users of A. This overlay plays an important role in a P2P recommender system. First, the performance of the recommendation depends on the structure of the overlay. Second, the costs and load balancing of the overlay management protocol depend on the topology of this similarity network.

To the best of our knowledge, the second role of the similarity overlay has not been addressed so far in the literature. Nevertheless it is an important issue, since the load generated by the overlay management process might correlate with the number of nodes that link to a given node as one of its most similar nodes. More precisely, the load of a node might correlate with its in-degree in the overlay network. Thus, if

Table 5.1: Basic statistics of datasets

	MovieLens	Jester	BookCrossing
# users	71,567	73,421	77,806
# items	10,681	100	185,974
size of train	9,301,274	3,695,834	397,011
sparsity	1.2168%	50.3376%	0.0027%
size of eval	698,780	440,526	36,660
eval/train	7.5127%	11.9195%	9.2340%
# items $\geq$	20	15	1
rate set	$1,\ldots,5$	$-10,\ldots,10$	$1,\ldots,10$
MAE(med)	0.93948	4.52645	2.43277

the in-degree distribution of the overlay network is extremely unbalanced (e.g. if it has a power-law distribution), some of the nodes can experience a load that is orders of magnitudes higher than the average. Thus, it is very important to consider the in-degree distribution of the overlay when planning a P2P recommender system, and examine the incurred loads on the individual nodes as a function of this distribution.

Figure 5.1 shows the in-degree distributions of the k nearest neighbor (kNN) overlay of each benchmark dataset. In this overlay each node has k directed outgoing edges to the k most similar nodes. As can be seen from the plots, the BookCrossing dataset has an almost power-law in-degree distribution, with many nodes having incoming links from almost every other node (note that the size of this dataset is around 77,806 users).

To see whether this might be a general property of high dimensional datasets, we need to consider some basic properties of high dimensional metric spaces. If we generate high dimensional uniform random datasets from the unit cube and construct their kNN graphs, we will find that most of the points lie on the convex hull of the dataset. These points are mostly situated at the same distance from each other. The nodes corresponding to these points have a mostly uniform and relatively small in-degree in the kNN graph. The very few points inside the convex hull are close to a huge number of points on the convex hull, and so have high in-degree.

These observations indicate that we have to explicitly take into account load balancing when building a recommender system in a fully distributed manner.

5.4 Algorithms

The algorithms we examine all rely on building and managing a user-similarity overlay. In the top level of the protocol hierarchy, they apply the same user-based CF algorithm for making recommendations, strictly using locally available information (that is, information about the neighbors in the overlay).

Since we focus on overlay management, we fix the recommender algorithm and not discuss it any further. As it was mentioned in the previous sections, for this we need an aggregation method and a user similarity metric. We selected the aggregation shown in (5.1), proposed in [98]. Our similarity metric is Cosine similarity, which achieved the best performance on our benchmarks. Note that the selected user similarity is of course known to the overlay management algorithm and is used to direct the overlay construction.

We also assume that the local views of the nodes contain not only the addresses of the neighbors, but also a descriptor for each neighbor, that contains ratings made by the corresponding user. This implies that computing recommendation scores do not load the network since all the necessary information is available locally. However, there is a drawback; namely the stored information is not up-to-date. As we will show later, this is not a serious problem since on the one hand, recommendation datasets are not extremely dynamic and, on the other hand, the descriptors are in fact refreshed rather frequently due to the management algorithms.

In sum, the task of overlay management is to build and maintain the best possible overlay for computing recommendation scores, by taking into account bandwidth us-

Algorithm 2 Random Nodes based Overlay Management

Parameters: k: the size of view; r: the number of randomly generated nodes

```
1. while true do
2.   samples ← getRandomPeers(r)
3.   for i = 1 to r do
4.     peer ← get(samples, i)
5.     peerDescriptor ← descriptor(peer)
6.     insert(view, peerDescriptor)
```

age at the nodes. We expect a minimal, uniform load from overlay management even when the in-degree distribution of the expected overlay graph is unbalanced.

5.4.1 BUDDYCAST based Recommendation

As we mentioned earlier we applied the BUDDYCAST overlay management protocol as a baseline method. Now we give a very brief overview of this algorithm and its numerous parameters; for details please see [95].

The algorithm maintains a number of lists containing node descriptors. The taste buddy list contains the most similar users (peers), all those who communicated with the node before. The recommendation for a peer is calculated based on this list.

The BUDDYCAST algorithm contains a mechanism for load balancing: a block list. Communication with a peer on the block list is not allowed. If a node communicates with another peer, it is put on the block list for four hours.

Finally, a node also maintains a candidate list, which contains close peers for potential communication, as well as a random list that contains random samples from the network. For overlay maintenance, each node periodically (in every 15 seconds by default) connects to the best node from the candidate list with probability α, and to a random list with probability $1 - \alpha$, and exchanges its buddy list with the selected peer.

5.4.2 kNN Graph from Random Samples

We assume that a node has a local view of size k that contains node descriptors. These will be used by the recommender algorithm.

In Algorithm 2 each node is initialized with k random samples from the network, and they iteratively approximate the kNN graph. The convergence is based on a random sampling process which generates r random nodes from the whole network in each iteration. These nodes are inserted into the view which is implemented as a bounded priority queue. The size of this queue is k and the priority is based on the similarity function provided by the recommender module.

Applying a priority queue here on the basis of similarities means that nodes remember the most similar nodes from the past iterations. This means that since random samples are taken from the entire network, each node will converge to its kNN view with positive probability.

Method GETRANDOMPEERS can be implemented, for example, using the NEWSCAST [58] protocol.

This algorithm does converge, as argued above, albeit very slowly. However, it is guaranteed to generate an almost completely uniform load since the only communication that takes place is performed by the underlying peer sampling implementation (NEWSCAST), which has this property.

5.4.3 kNN Graph by T-MAN

We can manage the overlay with the T-MAN algorithm as well [56]. This algorithm manages a view of size k, as in the random algorithm above. T-MAN periodically updates this view by first selecting a peer node to communicate with, then exchanging its view with the peer, and finally merging the two views and keeping the closest k descriptors. This is very similar to Algorithm 2, but instead of r random samples the update is performed using the k elements of the view of the selected peer.

In this chapter we examine the following methods for T-MAN which are employed as peer selection methods:

Global: This approach selects the node for communication from the whole network randomly. This can be done by using a NEWSCAST layer as it was described in the previous section. We expect this approach to distribute the load in the network uniformly since with this selection the incoming communication requests do not depend on the in-degree of the kNN graph at all.

View: In this approach the node for communication is selected from the view of the current node uniformly at random. The mechanism of this selection strategy is similar to the previous one, but the spectrum of the random selection is smaller since it is restricted to the view instead of the whole network.

Proportional: This approach also selects a node for view exchange from the view of the current node, but here we define a different probability distribution. This distribution is different for each node and it is reversely proportional to the value of a selection counter, which measures the load of the node in the previous time interval. The exact definition of the selection probability for a neighbor j of node i is

$$p_{i,j} = \frac{\frac{1}{sel_j+1}}{\sum_{k \in View_i} \frac{1}{sel_k+1}}, \tag{5.2}$$

where sel_k is the value of the selection counter of the kth neighbor. This information is stored in the node descriptors. The motivation for this selection method is to reduce the load on the nodes that have a high in-degree in the kNN graph, while maintaining the favorable convergence speed of the T-MAN algorithm.

Best: The strategy that selects the most similar node for communication without any restriction. We expect that this strategy converges the most aggressively to the perfect kNN graph, but at the same time it results in the most unbalanced load.

5.4.4 Randomness is Sometimes Better

Our experimental results (to be presented in Section 5.6) indicated that in certain cases it is actually *not* optimal to use the kNN view for recommendation. It appears to be the case that a more relaxed view can give better recommendation performance.

To test this hypothesis, we designed a randomization technique that is compatible with any of the algorithms above. The basic idea is that we introduce an additional parameter, $n \leq k$. The nodes still have a view of size k, and we still use the same recommender algorithm based on these k neighbors. However, we apply any of the algorithms above to construct a (k-n)NN overlay graph (not a kNN graph), and we fill the remaining n elements in the following way: we take $r \geq n$ random samples (not necessarily independent in each cycle) and we take the closest n nodes from this list. With $n = k$ we get the algorithms proposed in [10], and with $n = 0$ this modification has no effect, so we get the original algorithm for constructing the kNN graph.

5.5 System Model

We consider a set of nodes connected through a routed network. Each node has an address that is necessary and sufficient for sending a message to it. To actually communicate, a node has to know the address of the other node. This is achieved by maintaining a *partial view* (*view* for short) at each node that contains a set of node descriptors. Views can be interpreted as sets of edges between nodes, naturally defining a directed graph over the nodes that determines the topology of an *overlay network*.

Although the class of algorithms we discuss has been shown to tolerate unpredictable message delays and node failures well [56, 58], in this work we focus on load balancing and prediction performance, so we assume that messages are delivered reliably and without delay, and we assume that the nodes are stable.

Finally, we assume that all nodes have access to the peer sampling service [58] that returns random samples from the set of nodes in question. We will assume that these samples are indeed random. The results presented in [58] indicate that the peer sampling service has realistic implementations that provide high quality samples at a low cost.

5.6 Empirical Results

We implemented our protocols and performed our experiments in PeerSim [57, 79]. We performed a set of simulations of our algorithms with the following parameter value combinations: *view update* is random or T-MAN; *peer selection* for T-MAN is GLOBAL, VIEW, BEST or PROPORTIONAL; and the number of *random samples* is 20, 50, or 100 for random, and 0 or 100 for T-MAN.

The BUDDYCAST algorithm was implemented and executed with the following parameters: the size of the buddy list and the candidate list was 100, the size of the random list was 10, and α was 0.5. The size of the block list had to be restricted to be 100 as well, in order to be able to run our large scale simulations. The view size for the rest of the protocols was fixed at $k = 100$ in all experiments for practical reasons: this represents a tradeoff between a reasonably large k and the feasibility of large scale simulation.

In these simulations we observe the prediction performance in terms of the MAE measure and the distribution of the number of incoming messages per cycle at a node. Note that the number of outgoing messages is exactly one in each case.

Let us first discuss the effect of parameter r. This is a crucial parameter for random view update, while in the case of T-MAN the role of random samples is merely to help the algorithm to avoid local optima, and to guarantee convergence. Figure 5.2 shows the effect of r in the case of the MovieLens database. The effect of r on the other databases and for other settings is similar.

We can observe that in the case of a random view update, r simply is a multiplicative factor that determines the speed of convergence: twice as many samples per cycle result in a halving of the necessary cycles to achieve the same value. In the case of T-MAN, the version with random samples converges faster, while the generated load remains the same (not shown). Accordingly, in the following we discuss T-MAN algorithms only with $r = 100$, and random view update algorithms only with $r = 100$.

In Figure 5.3 we show the results of the experiments, where the MAE and the maximal load is illustrated. The maximal load is defined as the maximal number of incoming messages any node receives during the given cycle. The first interesting observation is that the load balancing property of the different algorithms shows a similar pattern over the three datasets, however, the convergence of the MAE is rather different (see also Table 5.1). In particular, in the case of the MovieLens and BookCrossing benchmarks the MAE reaches a minimum, after which it approaches the top-k based prediction from below, whereas we do not see this behavior in the much denser Jester database.

Indeed, the reason for this behavior lies in the fact that for the sparse datasets a larger k is a better choice, and our setting ($k = 100$) is actually far from optimal. In the initial cycles the view approximates a random sample from a larger k parameter. To verify this, we calculated the MAE of the predictions based on the algorithm described in Section 5.4.4. The results are shown in Figure 5.4 later on.

It is clear that for a small k it is actually better *not* to use the top k from the entire network; rather it is better to fill some of the views with the closest peers in a relatively small random sample from the network. Especially for the smallest k we examined ($k = 100$) this technique results in a significant improvement in the MAE compared to the recommendation based on the closest k peers in all datasets. This algorithm can easily be implemented, since we simply have to combine any of the convergent algorithms with an appropriate setting for k (such as $k = 50$) and use a peer sampling service to add to this list the best peers in a random sample of a given size.

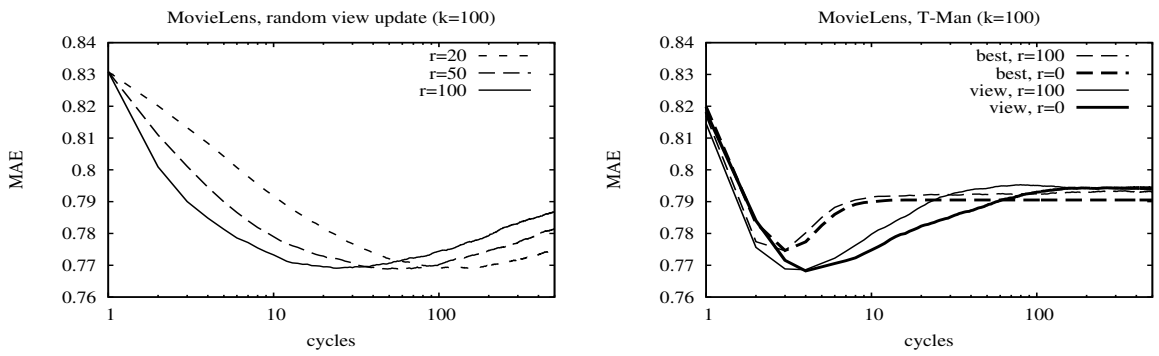

Figure 5.2: Effect of parameter r in a few settings

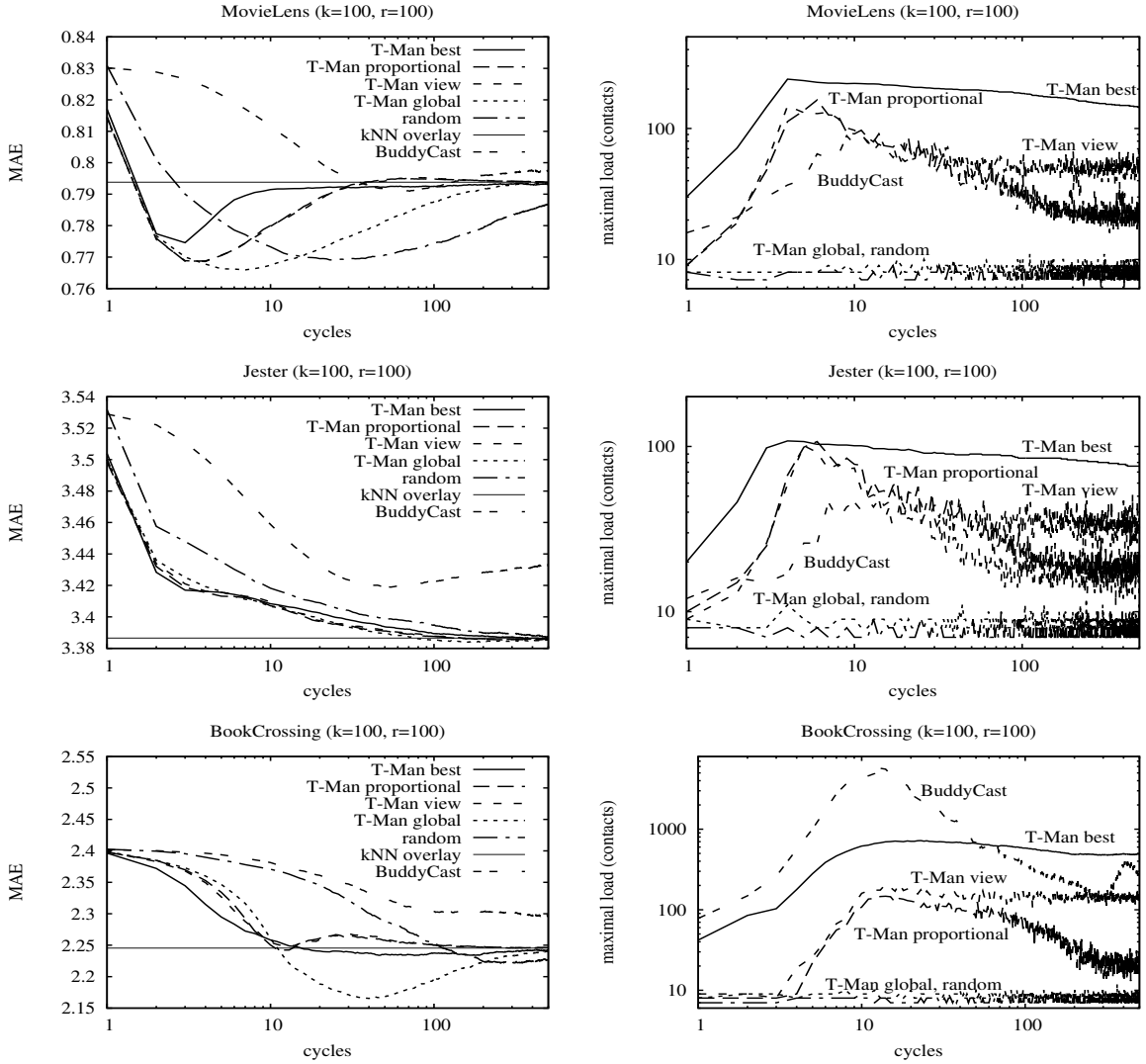

Figure 5.3: Experimental results. The scale of the plots on the right is logarithmic.

As a closely related note, the random view update algorithms can be "frozen" in the state of minimal MAE easily, without any extra communication, provided we know in advance the location (that is, the cycle number) of the minimum. Let us assume it is in cycle c. Then we can use, for a prediction at any point in time, the best k peers out of the union of $c \cdot r$ random samples collected in the previous c cycles, which is very similar to the approach taken in [10].

Clearly, the fastest convergence is shown by the T-MAN variants, but these result in unbalanced load at the same time. The PROPORTIONAL variant discussed in Section 5.4.3 reduces the maximal load, however, only when the topology has already converged. During the convergence phase, PROPORTIONAL behaves exactly like the variant VIEW.

Quite surprisingly, the best compromise between speed and load balancing seems to be GLOBAL, where the peer is selected completely at random by T-MAN. In many topologies, such as a 2-dimensional grid, a random peer possesses no useful information for another node that is far from it in the topology, so we can in fact expect to

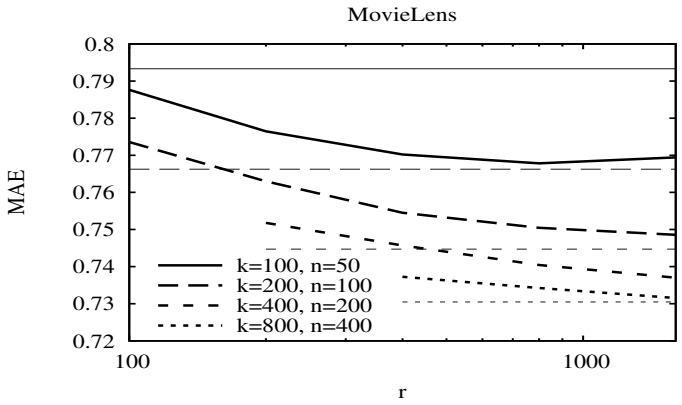

Figure 5.4: Effect of adding randomness to the view. Thin horizontal lines show the $n = 0$ case.

do worse than the random view update algorithm. However, in target graphs such as kNN graphs based on similarity metrics, a large proportion of the network shares useful information, namely the addresses of the nodes that are more central.

On such unbalanced graphs T-MAN GLOBAL is favorable, because it offers a faster convergence than a pure random search (in fact, it converges almost as fast as the more aggressive T-MAN variants), however, the load it generates over the network is completely identical to that of random search, and therefore the maximal load is very small: the maximum of N samples from a Poisson distribution with a mean of 1 (where N is the network size). In addition, the node with the maximal load is different in each cycle.

Finally, we can observe that on the BookCrossing database some algorithms, especially BuddyCast and T-MAN with BEST peer selection, result in an extremely unbalanced degree distribution (note the logarithmic scale of the plot). This correlates with the fact that the BookCrossing database has most unbalanced degree distribution (see Figure 5.1). Even though we have not optimized the parameters of BuddyCast, this result underlines our point that one has to pay attention to the in-degree distribution of the underlying kNN graph.

5.7 Conclusions

In this chapter we tackled the problem of the construction of similarity-based overlay networks with user-based collaborative filtering as an application. We pointed out that similarity-based overlays can have a very unbalanced degree distribution, and this fact might have a severe impact on the load balancing of some overlay management protocols. The main conclusion that we can draw is that in highly unbalanced overlays (that are rather frequent among similarity-based networks) the overlay construction converges reasonably fast even in the case of random updates; or, with T-MAN, uniform random peer selection from the network. At the same time, the traditional, aggressive peer selection strategies that have been proposed by other authors should be avoided because they result in a highly unbalanced load experienced by the nodes. In sum,

in this domain T-MAN with global selection is a good choice, because it has a fully uniform load distribution combined with an acceptable convergence speed, which is better than that of the random view update. However, care should be taken because this conclusion holds only in these unbalanced domains, and in fact this algorithm is guaranteed to perform extremely badly in large-diameter topologies.

Chapter 6

Conclusions

In summary, this deliverable describes how social interactions and collectives influence and overlap with opinion formation and how the resulting informational structures (in this case, fat-tailed degree distributions in overlay networks) need to be taken into account in the design of distributed algorithms. In particular, we presented a model based on propagation of influence in a social network and showed that when it is superimposed with unbiased expectations of individuals about respective items, a broad popularity distribution resembling popularity distributions seen in real systems typically arise. The model needs to be further improved to better mimic bipartite user-item data observed in real systems. We further presented a model for social cohesion. We argued that the social cohesion of a population should be expressed in terms of both structural and cognitive dimensions as a response to conflict increases. We are now looking forward to using the current on-line communication platforms (Twitter, Facebook, and alike) to empirically test theoretical models relying both on underlying bond topologies (who holds stable relations with whom) and on information dynamics (who is actually communicating with whom). In the final chapter, we addressed the problem of the construction of similarity-based overlay networks with user-based collaborative filtering as an application. Since similarity-based overlays can (and often do) have a very unbalanced degree distribution, aggressive peer selection strategies that have been proposed by other authors should be avoided because they result in a highly unbalanced load experienced by the nodes. This is not an issue for random updates and T-MAN. T-MAN with global selection thus emerges as a good choice in this domain because it has a fully uniform load distribution combined with an acceptable convergence speed, which is better than that of the random view update.

Bibliography

[1] R.P. Abelson. *Mathematical models of the distribution of attitudes under controversy*, pages 140–150. RinehartWinston, New York, U.S.A., 1964.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, pages 734–749, 2005.

[3] Gediminas Adomavicius and Er Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowledge and Data Engineering*, 17:734–749, 2005.

[4] R. Akavipat, L.-S. Wu, F. Menczer, and A.G. Maguitman. Emerging semantic communities in peer web search. In *Proc. Intl. workshop on Information retrieval in peer-to-peer networks (P2PIR'06)*, pages 1–8. ACM, 2006.

[5] R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholtz. Trust-based recommendation systems: an axiomatic approach. In *Proceeding of the 17th international conference on World Wide Web*, pages 199–208. ACM, 2008.

[6] E.W. Anderson and L.C. Salisbury. The formation of market-level expectations and its covariates. *Journal of Consumer Research*, 30(1):115–124, 2003.

[7] Rodrigo Araya. Multitudes y redes en la caida de milosevic. *REDES-Revista Hispana para el análisis de redes sociales*, 15(7), 2006.

[8] J. Arndt. Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research*, 4(3):291–295, 1967.

[9] Robert Axelrod. The dissemination of culture. *Journal of Conflict Resolution*, 41(2):203–226, 1997.

[10] Arno Bakker, Elth Ogston, and Maarten van Steen. Collaborative filtering using random neighbours in peer-to-peer networks. In *Proc. 1st ACM Intl. workshop on Complex networks meet information & knowledge management (CNIKM'09)*, pages 67–75. ACM, 2009.

[11] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[12] A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press New York, NY, USA, 2008.

[13] S. Bergamaschi, Francesco Guerra, and Barry Leiba. Information overload. *Internet Computing, IEEE*, 14(6):10–13, 2010.

[14] Danny Bickson, Dahlia Malkhi, and Lidong Zhou. Peer-to-Peer rating. In *Proc. 7th IEEE Intl. Conf. on Peer-to-Peer Computing, 2007. (P2P '07)*, pages 211–218. IEEE Computer Society, 2007.

[15] D. Billsus and M.J. Pazzani. Learning collaborative information filters. In *Proceedings of the Fifteenth International Conference on Machine Learning*, volume 54, page 48, 1998.

[16] Daniel Billsus and Michael J. Pazzani. Learning collaborative information filters. In *Proc. 15th Intl. Conf. on Machine Learning (ICML '98)*, pages 46–54. Morgan Kaufmann, 1998.

[17] A. Birukov, E. Blanzieri, and P. Giorgini. Implicit: An agent-based recommendation system for web search. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 618–624. ACM, 2005.

[18] M. Blattner. B-rank: A top N recommendation algorithm. In *Proceedings of The International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2010)*, volume 1, pages 337–341, Orlando, USA, 2010.

[19] M. Blattner and M. Medo. Recommendation systems in the scope of opinion formation: a model. *arXiv preprint arXiv:1206.3924*, 2012.

[20] Marián Boguñá, Romualdo Pastor-Satorras, Albert Díaz-Guilera, and Alex Arenas. Models of social networks based on social distance attachment. *Phys. Rev. E*, 70(5):056122, Nov 2004.

[21] J. Borge-Holthoefer, A. Rivero, I. García, E. Cauhé, A. Ferrer, D. Ferrer, D. Francos, D. Iñiguez, M.P. Pérez, G. Ruiz, et al. Structural and dynamical patterns on Online Social Networks: the Spanish May 15th movement as a case study. *PLoS One*, 6(8):e23883, 2011.

[22] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, San Francisco, CA, 1998. Morgan Kaufmann.

[23] P. Brusilovsky, A. Kobsa, and W. Nejdl. *The adaptive web: methods and strategies of web personalization*. Springer-Verlag New York Inc, 2007.

[24] G. Caron-Lormier, R.W. Humphry, D.A. Bohan, C. Hawes, and P. Thorbek. Asynchronous and synchronous updating in individual-based models. *ecological modelling*, 212(3-4):522–527, 2008.

[25] Sylvain Castagnos and Anne Boyer. Modeling preferences in a distributed recommender system. In *Proc. 11th Intl. Conf. on User Modeling (UM '07)*, pages 400–404. Springer-Verlag, 2007.

[26] Damon Centola, Juan Carlos González-Avella, Víctor M. Eguíluz, and Maxi San Miguel. Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution*, 51(6):905–929, 2007.

[27] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining content-based and collaborative filters in an online newspaper. In *Proc. ACM SIGIR 99, Workshop Recommender Systems: Algorithms and Evaluation*, 1999.

[28] J. Coleman. *Foundations of Social Theory*. Harvard University Press, Cambridge, U.S.A, 1990.

[29] M.H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69:118–212, 1974.

[30] H. Drachsler, T. Bogers, R. Vuorikari, K. Verbert, E. Duval, N. Manouselis, G. Beham, S. Lindstaedt, H. Stern, M. Friedrich, et al. Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1(2):2849–2858, 2010.

[31] A. Flache and M.W. Macy. What sustains cultural diversity and what undermines it? Axelrod and beyond. arXiv:physics/0604201v1, 2006.

[32] A. Flache and M.W. Macy. Local Convergence and Global Diversity: The Robustness of Cultural Homophily. arXiv:physics/0701333v1, 2007.

[33] F. Fouss, A. Pirotte, J.M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):355–369, 2007.

[34] F. Fouss, L. Yen, A. Pirotte, and M. Saerens. An experimental investigation of graph kernels on a collaborative recommendation task. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 863–868. IEEE, 2007.

[35] P. Garbacki, A. Iosup, J. Doumen, J. Roozenburg, Y. Yuan, Ten M. Brinke, L. Musat, F. Zindel, F. van der Werf, M. Meulpolder, and Others. Tribler protocol specification.

[36] Pawe Garbacki, Dick H. J. Epema, and Maarten van Steen. A two-level semantic caching scheme for super-peer networks. In *Proc. 10th Intl. Workshop on Web Content Caching and Distribution (WCW'05)*, pages 47–55. IEEE Computer Society, 2005.

[37] W. Geyer, J. Freyne, B. Mobasher, S.S. Anand, and C. Dugan. 2nd workshop on recommender systems and the social web. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 379–380. ACM, 2010.

[38] A. Giddens. *The Constitution of Society. Outline of the Theory of Structuration*. Polity, Cambridge, 1984.

[39] J.P. Gleeson. High-accuracy approximation of binary-state dynamics on networks. *Physical Review Letters*, 107(6):68701, 2011.

[40] D. Goldberg, B.M. D. Nichols, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[41] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[42] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The dynamics of protest recruitment through an online network. *Scientific Reports*, 1:197, 2011.

[43] N. Good, J.B. Schafer, J.A. Konstan, A. Brochers, B. Sarwar, J.L. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations. In *Proc. Conf. Am. Assoc. Artificial Intelligence (AAAI-99)*, pages 439–446, USA, 1999.

[44] R. V. Gould. Multiple networks and mobilization in the paris commune, 1871. *American Sociological Review*, 56:716–729, 1991.

[45] R. V. Gould. Collective action and network structure. *American Sociological Review*, 58(2):182–196, 1993.

[46] R. V. Gould. *Insurgent Identities: Class, Community, and Protest in Paris from 1848 to the Commune*. University of Chicago Press, Chicago, U.S.A, 1995.

[47] T. Gross and B. Blassius. Adaptive coevolutionary networks: A review. *J. R. Soc. Interfac.*, 5(20):259–271, 2008.

[48] I. Guy, A. Jaimes, P. Agulló, P. Moore, P. Nandy, C. Nastar, and H. Schinzel. Will recommenders kill search?: recommender systems-an industry perspective. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 7–12. ACM, 2010.

[49] Peng Han, Bo Xie, Fan Yang, and Ruimin Shen. A scalable P2P recommender system based on distributed collaborative filtering. *Expert Systems with Applications*, 27(2):203–210, 2004.

[50] T. Hennig-Thurau, K.P. Gwinner, G. Walsh, and D.D. Gremler. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1):38–52, 2004.

[51] J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[52] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proc. 22nd annual Intl. ACM SIGIR Conf. on Research and development in information retrieval (SIGIR '99)*, pages 230–237. ACM, 1999.

[53] P Holme and M E.J Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74:056108, 2006.

[54] M.O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, U.S.A, 2008.

[55] D. Jannach, W. Geyer, J. Freyne, S.S. Anand, C. Dugan, B. Mobasher, and A. Kobsa. Recommender Systems & the Social Web. In *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009*. ACM, 2009.

[56] Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. T-Man: Gossip-based fast overlay topology construction. *Computer Networks*, 53(13):2321–2339, 2009.

[57] Márk Jelasity, Alberto Montresor, Gian Paolo Jesi, and Spyros Voulgaris. The Peersim simulator. `http://peersim.sf.net`.

[58] Márk Jelasity, Spyros Voulgaris, Rachid Guerraoui, Anne-Marie Kermarrec, and Maarten van Steen. Gossip-based peer sampling. *ACM Trans. on Computer Systems*, 25(3):8, 2007.

[59] Anne-Marie Kermarrec. Challenges in personalizing and decentralizing the web: An overview of GOSSPLE. In *Proc. 11th Intl. Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2009)*, volume 5873 of *LNCS*, pages 1–16. Springer, 2009.

[60] K.Goldberg, T. Roeder, D Gupta, and C. Perkins. Eigentaste: a constant time collaborative filtering algorithm. *Information Retrieval*, 4:133–151, 2001.

[61] Y.A. Kim and J. Srivastava. Impact of social influence in e-commerce decision making. *Proceedings of the ninth international conference on Electronic commerce (ICEC)*, pages 293–302, 2007.

[62] V. et al Klemm, K. Eguiluz. Global culture: A noise-induced transition in finite systems. *Physical Review E*, 67:045101R, 2003.

[63] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, and L.R. Gordon. Grouplens: Applying collaborative filtering to usenet news. *Comm. ACM*, 40(3):77–87, 1997.

[64] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[65] Neil D. Lawrence and Raquel Urtasun. Non-linear matrix factorization with gaussian processes. In *Proc. 26th Annual Intl. Conf. on Machine Learning (ICML '09)*, pages 601–608. ACM, 2009.

[66] D. Lazer. The co-evolution of individual and network. *Journal of Mathematical Sociology*, 25:69–108, 2001.

[67] H Liao, G Cimini, and M Medo. *ISMIS 2012/Lecture Notes in Artificial Intelligence 7661*, page 421. Springer-Verlag, New York, U.S.A., 2012.

[68] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[69] S. Lozano, J. Borge-Holthoefer, and A. Arenas. Emerging cohesion and individualization in collective action: A co-evolutive approach. *Advances in Complex Systems*, 15(supp01), 2012.

[70] A. Mäs, M. Flache and James A. Kitts. Cultural Integration and Differentiation in Groups and Organizations. In review.

[71] Michael Mäs, Andreas Flache, and Dirk Helbing. Individualization as driving force of clustering phenomena in humans. *PLoS Comput Biol*, 6(10):e1000959, 10 2010.

[72] M.F. Mason, R. Dyer, and M.I. Norton. Neural mechanisms of social influence. *Organizational Behavior and Human Decision Processes*, 110(2):152–159, 2009.

[73] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 492–508, 2004.

[74] P. Massa and B. Bhattacharjee. Using trust in recommender systems: an experimental analysis. *Trust Management*, pages 221–235, 2004.

[75] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:pp. 415–444, 2001.

[76] P. Melville, R.J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Proc. 18h Nat'l Conf. ARtificial Intelligence*, 2002.

[77] Q. Michard and J.P. Bouchaud. Theory of collective opinion shifts: from smooth trends to abrupt swings. *The European Physical Journal B-Condensed Matter and Complex Systems*, 47(1):151–159, 2005.

[78] B.J. Mirza, B.J. Keller, and N. Ramakrishnan. Studying recommendation algorithms by graph analysis. *Journal of Intelligent Information Systems*, 20(2):131–160, 2003.

[79] Alberto Montresor and Márk Jelasity. Peersim: A scalable P2P simulator. In *Proc. Ninth IEEE Intl. Conf. on Peer-to-Peer Computing (P2P 2009)*, pages 99–100. IEEE, 2009. extended abstract.

[80] J Moody and D.R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):103–127, 2003.

[81] R.F. Murphy. Intergroup hostility and social cohesion. *American Anthropologist*, 59(6):1018–1035, 1957.

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

[83] M.E.J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

[84] M. O'Connor and J. Herlocker. Clustering items for collaborative filtering. *Workshop on Recommender Systems at 22nd ACM SIGIR*, 1999.

[85] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 167–174. ACM, 2005.

[86] Karl-Dieter Opp and Christiane Gern. Dissident groups, personal networks, and spontaneous cooperation: The east german revolution of 1989. *American Sociological Review*, 58(5):659–680, 1993.

[87] R. Ormándi, I. Hegedűs, and M. Jelasity. Overlay management for fully distributed user-based collaborative filtering. *Euro-Par 2010-Parallel Processing*, pages 446–457, 2010.

[88] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Asynchronous peer-to-peer data mining with stochastic gradient descent. In *17th International European Conference on Parallel and Distributed Computing (Euro-Par 2011)*, volume 6852 of *Lecture Notes in Computer Science*, pages 528–540. Springer-Verlag, 2011.

[89] Róbert Ormándi, István Hegedüs, and Márk Jelasity. Gossip learning with linear models on fully distributed data. *Concurrency and Computation: Practice and Experience*, pages n/a–n/a, 2012.

[90] Yoon-Joo Park and Alexander Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proc. 2008 ACM Conf. on Recommender systems (RecSys '08)*, pages 11–18. ACM, 2008.

[91] M. Pazzani and D. Billsus. Content-based recommendation systems. *Lecture Notes Computer Science*, 4321:325–341, 2007.

[92] M. Perc and A. Szolnoki. Coevolutionary games - a mini review. *BioSystems*, 99:109–125, 2010.

[93] M Pineda, R Toral, and E Hernández-García. Noisy continuous-opinion dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(08):P08001, 2009.

[94] Georgios Pitsilis and Lindsay Marshall. A trust-enabled P2P recommender system. In *Proc. 15th IEEE Intl. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE '06)*, pages 59–64, 2006.

[95] J.A. Pouwelse, J. Yang, M. Meulpolder, D.H.J. Epema, and H.J. Sips. Buddycast: an operational peer-to-peer epidemic protocol stack. In *Proc. 14th Annual Conf. of the Advanced School for Computing and Imaging*, pages 200–205. ASCI, 2008.

[96] Hegselmann R. and Krause U. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation (JASSS)*, 5(3), 2002.

[97] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netwews. In *Proc. Computer Supported Cooberative Work Conf.*, 1994.

[98] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proc. 1994 ACM Conf. on Computer supported cooperative work (CSCW '94)*, pages 175–186. ACM, 1994.

[99] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, 40:56–58, March 1997.

[100] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[101] M.L. Richins and T. Root-Shaffer. THE ROLE OF EVOLVEMENT AND OPINION LEADERSHIP IN CONSUMER WORD-OF-MOUTH: AN IMPLICIT MODEL MADE EXPLICIT. *Advances in consumer research*, 15:32–36, 1988.

[102] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM Press.

[103] C.E.G. Snijders, T.A.B. Steglich and M. Pearson. Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, 40:329–393, 2010.

[104] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, pages 675–684. ACM, 2004.

[105] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Scalable collaborative filtering approaches for large recommender systems. *Journal of Machine Learning Research*, 10:623–656, 2009.

[106] L. et al. Toivonen, R. Kovanen. A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks*, 39:240–254, 2009.

[107] Amund Tveit. Peer-to-peer based recommendations for mobile commerce. In *Proc. 1st Intl. workshop on Mobile commerce (WMC '01)*, pages 26–29. ACM, 2001.

[108] Federico Vazquez, Víctor M. Eguíluz, and Maxi San Miguel. Generic absorbing transition in coevolution dynamics. *Phys. Rev. Lett.*, 100:108702, Mar 2008.

[109] Spyros Voulgaris and Maarten van Steen. Epidemic-style management of semantic overlays for content-based searching. In *Proc. Euro-Par*, number 3648 in LNCS, pages 1143–1152. Springer, 2005.

[110] F.E. Walter, S. Battiston, and F. Schweitzer. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems*, 16(1):57–74, 2008.

[111] Jun Wang, Arjen P. de Vries, and Marcel J. T. Reinders. Unified relevance models for rating prediction in collaborative filtering. *ACM Trans. on Information Systems (TOIS)*, 26(3):1–42, 2008.

[112] G.I. Webb, M.J. Pazzani, and D. Billsus. Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1):19–29, 2001.

[113] Douglas R. White and Frank Harary. The cohesiveness of blocks in social networks: Node connectivity and conditional density. *Sociological Methodology*, 31:305–390, 2001.

[114] W.H. Whyte Jr. The web of word of mouth, 1954.

[115] T. Zhang and V.S. Iyengar. Recommender systems using linear classifiers. *The Journal of Machine Learning Research*, 2:334, 2002.

[116] Y.C. Zhang, M. Blattner, and Y.K. Yu. Heat conduction process on community networks as a recommendation model. *Physical review letters*, 99(15):154301, 2007.

[117] T. Zhou, J. Ren, M. Medo, and Y.C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):46115, 2007.

[118] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proc. 14th Intl. Conf. on WWW*, pages 22–32. ACM, 2005.

[119] D. Zwillinger and S. Kokoska. *CRC standard probability and statistics tables and formulae*. CRC, 2000.