
Master thesis
60 credits

Per Kristian Haakonsen

October 29th 2007
When a society is rich, its people don’t need to work with their hands; they can devote themselves to activities of spirit. We have more and more universities and more and more students. If students are going to earn degrees, they’ve got to come up with dissertation topics. And since dissertations can be written about everything under the sun, the number of topics is infinite. Sheets of paper covered with words pile up in archives sadder than cemeteries, because no one ever visits them, not even on All Souls’ Day. Culture is perishing in overproduction, in an avalanche of words, in the madness of quantity.

- Milan Kundera, The Unbearable Lightness of Being
Abstract
The usage of collaborative filtering mechanisms has become widespread on the web. But how reliable are such recommendation systems as a basis for making out choices? This thesis starts by compiling a list of known weaknesses of rating scales and written reviews, based on existing work within the field of social navigation and a few other disciplines. Four Norwegian web sites, all implementing collaborative filtering techniques, are then analyzed in relation to the compilation of known weaknesses. Based on this analysis, I conclude, with some limitations, that the four cases are highly unreliable as recommendation systems. This conclusion can be extended beyond the four cases studied, since it can make users more critical to the usage of recommendations found on the web, and enable designers to construct more reliable recommendation systems.

Keywords: Human-computer interaction, HCI, social navigation, collaborative filtering, CF, recommendation systems, web, Internet.
## Contents

1. Introduction .........................................................................................................................6  
   1.1 Motivation and Background .........................................................................................6  
   1.2 Research Question ......................................................................................................7  
   1.3 A Short Overview .......................................................................................................9  
   1.4 Existing Research .....................................................................................................9  
2. Theoretical Framework .....................................................................................................12  
   2.1 Human-Computer Interaction ..................................................................................12  
      2.1.1 A Short History .................................................................................................12  
      2.1.2 An Attempt to Define .......................................................................................13  
      2.1.3 An Interdisciplinary Field ..................................................................................14  
      2.1.4 How is HCI studied? .........................................................................................15  
      2.1.5 Why is HCI important? .....................................................................................17  
   2.2 Social Navigation ......................................................................................................19  
      2.2.1 Semantic, Spatial and Social Navigation .........................................................22  
      2.2.2 Space and Place ...............................................................................................24  
      2.2.3 Direct, Indirect and Social Texture ....................................................................26  
      2.2.4 Social Navigation and HCI .............................................................................26  
   2.3 Social Navigation Rating Techniques .......................................................................27  
      2.3.1 Collaborative Filtering ......................................................................................28  
      2.3.2 Written Reviews ..............................................................................................33  
   2.4 Challenges of Social Navigation Rating Techniques .................................................34  
      2.4.1 Algorithm .........................................................................................................35  
      2.4.2 Shilling Attacks .................................................................................................36  
      2.4.3 Barrier to Entry .................................................................................................38  
      2.4.4 Conformity ........................................................................................................41  
      2.4.5 Rating Scales ......................................................................................................42  
      2.4.6 Confidence ..........................................................................................................45  
      2.4.7 Fading ................................................................................................................47  
      2.4.8 Privacy ..............................................................................................................48  
   2.5 Challenges of User Reviews ......................................................................................49  
3. Method .................................................................................................................................51  
   3.1 Method Theory ..........................................................................................................51  
      3.1.1 Collecting Data .................................................................................................51  
      3.1.2 Analyzing the Data ............................................................................................52  
   3.2 Selecting the Cases ....................................................................................................54  
   3.3 Studying the Cases ....................................................................................................55  
      3.3.1 Shilling attacks .................................................................................................56  
      3.3.2 Barrier to Entry .................................................................................................57  
      3.3.3 Conformity ........................................................................................................57  
      3.3.4 Rating Scales ......................................................................................................58  
      3.3.5 Confidence .........................................................................................................59  
      3.3.6 Fading ................................................................................................................59  
      3.3.7 Written Reviews ...............................................................................................59  
4. Introducing the Cases ........................................................................................................62  
   4.1 Akam ..........................................................................................................................62  
   4.2 DinPris .........................................................................................................................64  
   4.3 Kelkoo .........................................................................................................................65


1. Introduction

Only two things are infinite, the universe and human stupidity, and I'm not sure about the former.

- A. Einstein

1.1 Motivation and Background

The author was recently in London. As any good tourist, a visit to Madam Tussaud’s was on the menu. The basis for this choice was numerous stories from other people – a selection made through social navigation. After examining a queue that was probably visible from the moon and pondering the guidebooks description of Madam Tussaud’s as …anything more than a staggeringly overpriced tourist trap, I decided not to do what others had done before me. The purpose behind this story is not to scare you from visiting Madam Tussaud’s, but rather to illustrate how social navigation failed. Freyne writes that social navigation harvest[s] and harness[es] community wisdom (Freyne et al., 2007, p. 52), but the example above rather illustrates an example of community stupidity. This was initially what made me want to take a critical look at some of the techniques within the field of social navigation.

Web sites which employ social navigation techniques, of which recommendation systems are one technique, have become widespread on the Internet. I prime example occurred as this thesis was being written. Within a few months in the Winter/Spring of 2007, the usage of the social utility Facebook (www.facebook.com) exploded in Norway, to the point where it made mainstream-media headlines, clogged the author’s inbox with invitations to join and the phrase “are you on Facebook yet?” occurred in nearly all conversations. Facebook, in addition to being a social utility for keeping in touch with people around you, can act as a recommendation system in that you can share links to whatever you may find interesting on the web. The concept behind recommendation systems is, however, far from new. Most ancient and medieval communities relied on word of mouth as the primary enabler of economical and social activity (Dellarocas, 2003, p. 1410), and word of mouth is a physical-world social navigation technique. One can claim that receiving recommendations, tips, gossip, advice and the like has existed as long as humans. It is a part of our society. How
many times haven’t you seen a movie, listened to a song, read a book, etc. based on a friend’s recommendation?

The Internet provides new opportunities for this age-old phenomenon. One can point at two central new characteristics. First, anyone can convey their opinion on about anything. The number of web sites which offer some sort of recommendation system, and the range of topics which these sites cover, gives any individual with access to the Internet the opportunity of expressing one’s opinion on the subject in question. The second difference concerns reach. Conveying recommendations to the masses used to be a privilege of the few, such as book reviewers in newspapers or on the radio. The Internet has created the opportunity for sharing one’s views and opinions to the whole world with a few clicks of the mouse. By now, anyone who has access to the Internet can share their opinions, views and the like on about any topic with other individuals from all over the globe.

With the access to a vast pool of other users’ opinions available at one’s fingertips, people increasingly rely on opinions posted on these systems to make a variety of decisions ranging from which movie to watch next to which stocks to invest in (Dellarocas 2003 p. 1408-1409; Godes and Mayzlin, 2004, p. 546). As usage increases, the question of reliability grows increasingly important. How reliable are the systems themselves? How reliable is this kind of user-created content?

1.2 Research Question
An important aspect of social navigation theory is that social navigation can help turning a space into a place. Through the implementation of the techniques found in social navigation, a web site can be transformed from a lonely space in which users are unaware of each other into a place where users become aware of each other, a place where users can contribute to site content and get a sense of ownership and participation. But what kind of place? This work will focus on web-based product recommendation sites and seek to give an answer to the following question:

*How reliable are web-based recommender systems as an aid for making our choices?*
This question poses new questions. How do you measure reliability? In order to further specify this question, I have divided it into three parts:

1. Through the study of social navigation theory and existing research done on social navigation systems, a set of known weaknesses can be compiled. Through an analysis of existing web-based recommendation systems, how exposed are these to the known weaknesses?
2. By analyzing web-based recommendation systems, are there any weaknesses not identified through question 1?
3. By analyzing web-based recommendation systems, are there any strengths not identified through question 1?

Focusing on the weaknesses of web-based recommender systems can reveal how reliable these systems actually are. Additionally, knowledge of weaknesses can help to unveil how such systems can be improved in order to make them more reliable. On the other hand, too much focus on weaknesses could create a falsely negative picture. Hence, when compiling the list of known weaknesses (sub-question 1), I will also look into known solutions to these weaknesses and investigate whether these solutions are implemented in the recommendation systems studied. In addition, in sub-question 3 I will look at strengths not yet discovered. By analyzing the recommendation systems from both angles, I hope to create a balanced analysis of the recommender systems studied.

As noted, focusing on weaknesses can help discover how recommendation systems can be more reliable. Hence, in the final conclusion, I will analyze which implications the results found have for the user, social navigation theory and the designer of these systems. In this way, I hope to create a two-fold conclusion, focusing not only on how reliable the recommendation systems studied are, but also which general implications these results have for the three parties above.

There are some limitations to the approach chosen. The number of recommendation systems found on the web is immense, and hence a selection has to be made. This naturally has the consequence that the evaluation of reliability applies only to the selection of recommendation systems studied. On the other hand, implications for the user, social navigation theory and the
designer are hopefully more general and can be applied across a wider range of recommendation systems.

1.3 A Short Overview
The general approach used in this thesis is as follows. First, an overview of existing known weaknesses of social navigation rating techniques are collected and categorized. The sources of this collection of weaknesses are based on social navigation theory and other individuals’ work, experiments and research. Second, a selection of four Norwegian recommendation web sites is analyzed in relation to the known set of weaknesses. Additionally, a collection of newly found strengths and weaknesses will be presented. Based on this analysis I will assess the reliability of the four recommendation sites and hence how trustworthy they are as a basis for making our choices. Finally, general implications for the user, social navigation theory and the designer will be presented.

The chapters will proceed as follows. Chapter two is the theoretical introduction. At first I will introduce Human-Computer Interaction (HCI) and then move on to introduce important aspects of social navigation. After this I will focus the theoretical presentation on recommendation systems, and finish the chapter by presenting an overview of known challenges and weaknesses of these systems. Chapter three describes the methods used in this work. First, the method theory is presented. In the two following sections, I describe how the four cases were selected and finally which methods were applied when studying the cases. The methods are derived from the known list of weaknesses and challenges identified in chapter two. The fourth chapter gives a brief introduction to the four cases studied, and in chapter five the research findings are presented. In chapter six we’ll reach the peak of excitement as the results are analyzed and discussed in relation to the theoretical framework. Each known challenge and weakness identified in the theory section and analyzed in relation to my cases will be discussed separately. Finally, in chapter seven, we’ll reach the end of the journey as the conclusions of this work are presented.

1.4 Existing Research
This section will present existing research into social navigation recommendation systems. The works presented here have similarities to the approach taken in this thesis, however they are not entirely alike. I have found no previous research very similar to this thesis, nor any
work done on the cases chosen for study in this thesis. Before we start, it is important to notice that the analysis of known weaknesses of recommendation systems presented in the theory section relies heavily on existing research. Thus, as this section presents existing work on the topic at hand, further presentation of existing research is found in the theory chapter.

In their paper “The Effect of Word of Mouth on Sales: Online Book Reviews” Chevalier and Mayzlin studies the effect of customer reviews on sales of books on the online book stores Amazon.com and BarnesandNoble.com (Chevalier and Mayzlin, 2003). Their approach is to gather sales rankings and user ratings from a random set of books and analyze these. They point at three central findings. First, reviews are overwhelmingly positive on both sites. Second, improvement in a book’s reviews leads to increase in relative sales. Third, the impact of a 1-star review is greater than an impact of a 5-star review. They conclude that customer reviews in these cases have an important impact on customer behaviour.

The second paper continues where Chevalier and Mayzlin left off. In their paper “Can Online Reviews Reveal a Product’s True Quality?”, Hu, Pavlou and Zhang looks into whether online reviews can reveal the true quality of a product (Hu, Pavlou and Zhang, 2006). Their approach is to study the distribution of user reviews of books, DVDs and videos on Amazon.com. They conclude that about half of the product sample does not necessarily reveal a product’s true quality and may hence provide misleading recommendations.

In the last paper “The Digitalization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms”, Dellarocas look into the feedback mechanism of the online auction site eBay (Dellarocas, 2003). The author does not himself study the site, but presents an overview of previous research done on the eBay feedback mechanisms. There are three important conclusions relevant to this work. First, feedback is overwhelmingly positive. Second, feedback seems to affect both prices and the probability of sale, though there are studies pointing in the other direction. Third, among the different feedback information eBay displays about its users, the overall number of positive and negative ratings are most influential in affecting buyer behaviour.

My research differs from the work described above in that I employ a somewhat broader approach. The above research focuses mainly on ratings given by the sites’ users. My work, in addition to looking at these ratings, looks at a wider array of the rating systems’ properties
such as strengths and weaknesses found within the recommendation systems themselves and usefulness of written reviews created by other users.

As an ending remark, it is interesting to note that none of the work presented above is done in a social-navigation framework, whereas they clearly work on typical social navigation topics. Of the above works two were found within management literature and one within literature on electronic commerce.
2. Theoretical Framework

*Man-computer symbiosis is an expected development in cooperative interaction between men and electronic computers. It will involve very close coupling between the human and the electronic members of the partnership. The main aims are 1) to let computers facilitate formulative thinking as they now facilitate the solution of formulated problems, and 2) to enable men and computers to cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs.*

(Licklieder, 1960, p. 4)

2.1 Human-Computer Interaction

This section will give an overview of the field Human Computer Interaction (HCI) of which social navigation is considered a sub-field. I will start by giving an historical overview of the field of HCI and how it evolved through time. I will then move on and attempt to define HCI and explain why it is regarded as a multidisciplinary field. Finally, I will introduce how HCI is applied in studies and why this field is important.

2.1.1 A Short History

The field of HCI can be traced back to the 1940s. The strive to develop more effective weapons systems, fuelled by the Second World War and combined with the increasing complexity of mechanical machines, which required the utmost effort of the human mind to operate, caused the emergence of applied research on humans and machine interfaces (Dix et al., 2004, p. 3; Butler, Jacob and John, 1999, p. 100). The work within human-machine interaction was primarily focused around ergonomics, concerned mainly with the physical characteristics of machines and systems, and how these affected user performance (Dix et al., 2004, p. 3). This led to the formation of the Ergonomics Research Society in 1949. Papers on the subject of HCI began to appear in the 1960s (Nickerson and Landauer, 1997, p. 11).

With the emergence of computers, the field of HCI evolved and became more focused on physical, psychological and theoretical aspects of human-computer interaction. The term
human-computer interaction became widely used in the 1980s and 1982 saw the first conference on human factors in computing systems held in Gaithersburg, Maryland (Dix et al., 2004, p. 3; Karat and Karat, 2003, p. 532). The journal *Human-Computer Interaction* was founded in 1985 (Nickerson and Landauer, 1997, p. 11). When computers first appeared, users had, for the most part, technical experience or at least some interest in computers and computer science. Programming skills were common and users were willing to invest time in learning how these machines worked. Developers were designing systems for their own use or for other technically proficient users. (Grudin, 1991, p. 293). Today, this landscape has changed considerably. As the use of computers has spread into all corners of society, the users of the technology have become a heterogeneous group. The technically experienced user still exists, but a large group of today’s users do not possess this knowledge or the interest in obtaining knowledge within this field at all. The use of computers is for many simply a tool through which they can achieve a goal. This transition caused a shift of focus within the field of HCI. At the outset, focus within HCI was on the specialist, but as the usage of computers spread, focus within HCI shifted to examining how technology impacts us all (Karat and Karat, 2003, p. 533). This shift constituted a great challenge for the field of HCI. As users of computer technology today are a heterogeneous group, varying in age, computer skills, cultural background and so on, creating user-friendly environments becomes increasingly difficult. This makes HCI more important than ever. A heterogeneous base of users requires a lot more focus on creating usable and useful environments in order to capture the needs of the diverse pool of users.

The field human-computer interaction was initially born as the result of the ever-increasing complexity in mechanical machines in the first half of the 20th century. As science developed the computer, HCI became a part of this field as well. Initially, focus within HCI was on the expert user, but as the usage of computers spread to all strata of society, focus within HCI shifted from the expert user to how technology impacts us all.

### 2.1.2 An Attempt to Define

Today, HCI has become a broad, interdisciplinary field, and hence there is no general unified theory within HCI. However, there are attempts to define the field in rather broad terms. Dix et al. describes HCI as *involv[ing] the design, implementation and evaluation of interactive systems in the context of the user’s task and work* (Dix et al., 2004, p. 4). They further
elaborate this statement, and explain *user, computer* and *interaction*. The user is whoever is trying to get the job done, be it a single person, a group or a sequence of users. The computer includes any computerized system, ranging from desktops to mainframes. The last participant, interaction, is defined as any communication between user and computer. Communication can again be categorized into *direct* and *indirect*. Direct communication involves dialog with feedback, such as using a browser to view Web pages. Examples of indirect communication could be batch jobs or intelligent sensors monitoring an environment.

Out of this definition, a fourth area of focus emerges, namely usability. In order for a computerized system to enable users to interact, the system should be easy to use. Dix et al. divides usability into three major parts, *useful, usable* and *used* (Dix et al., 2004, p. 5). A system is useful when it enables the user to accomplish what is required. Second, a system must be usable, enabling the user to accomplish a task easily and naturally. Last, a system must be used. It must make people want to use it, be attractive and engaging. It should be noted that the terms *useful, usable* and *used* are also known under the names *effectiveness, efficiency* and *satisfaction* (Karat and Karat, 2003, p. 535).

There is no general unified theory within HCI. However, in order to broadly identify the field, HCI has its focus on the human and the computer, the interaction between these and the usability of this interaction.

**2.1.3 An Interdisciplinary Field**

As previously mentioned, HCI is an interdisciplinary field. In this section I will give a short step-by-step introduction into the most important fields related to HCI, including some examples of what HCI derives from these sciences.

- First we have psychology and cognitive sciences. This field concerns subjects like the capabilities of human memory, motor skills, how people communicate with each other and social dynamics (Dix et al., 2004, p. 4). As an example, the 7±2 rule\(^1\) is derived from this field.
- Ergonomics is involved in order to understand the user’s physical capabilities (Baecker et al., 1995a, p. 40-41). In the world of mobile phone design for instance,

---

\(^1\) The 7±2 rule states that the average human is not capable of remembering more than 7±2 chunks of information.
ergonomics become important because small units require small buttons, which often are troublesome because you can't reduce the size of a human hand.

- Sociology is involved in order to understand the wider context of the interaction (Dix et al., 2004, p. 4). Technology changes society, and by better understanding how, one can include this knowledge in the design of computerized systems.
- Graphical design and layout is of course of great importance, since it’s the main channel through which communication between human and computer is done (Baecker et al., 1995a, p. 38-39).
- Most computerized systems require a manual, and thus the field of technical writing steps into the world of HCI (Dix et al., 2004, p. 4).
- In the development of commercial computer systems, getting someone to actually buy the system is of course vital. Hence, business becomes a part of HCI (Dix et al., 2004, p. 4).
- The field of artificial intelligence can help develop ideas concerning how to automate more work or make computers that behave more like intelligent assistants (Baecker et al., 1995a, p. 43-44).
- And finally, computer science and engineering is naturally involved, as the tool through which the technology is built (Dix et al., 2004, p. 4).

Clearly, there is a multitude of fields related to the work done within HCI, and theoretically more fields could be added if they can provide any useful contribution to the study of the relationship between humans and computers.

2.1.4 How is HCI studied?

At this point, HCI might seem a bit confusing. It is a broad, interdisciplinary field, focusing on the user in context. But how is HCI studied in practice? This section will introduce an overview of some of the most important examples of how HCI is studied. These six methods, described below, are divided into categories. However, the borders between the methods are sometimes blurry, and the usage of several of these methods, or parts of the methods, simultaneously in the development process of a computerized system can be a very beneficial approach.

*Task analysis* is a collection of formal and informal techniques which can be applied in order to find out which goals the system is to fulfil, and how technology can be used to reach these
goals (Nickerson and Landauer, 1997, p. 16). One common method is for the developer to visit the workplace where the system is to be used, look at how people do their job and talk to employees in order to create a clearer understanding of the system’s role.

Not too far away from task analysis we find consultation of potential users. This approach is often called “Scandinavian School” (Bødker, Grønbæk and Kyng, 1995, pp. 215-217; Grudin, 1991, p. 293). The idea behind this approach is that it is very hard to understand potential users unless you understand their situation, for instance the processes and culture of a workplace. The system developers do not initially understand the user in the context of their workplace, and should hence go into the workplace, understand the situation of the workers and include them in the design process. A second important focal area in this approach is to understand how the introduction of new technology in a workplace can change how people work and their work-situation. For instance, the introduction of a new system could make a specific position redundant; hence an employee might be in danger of becoming unemployed.

Our next approach is called formative design evaluation. The purpose behind this approach is to identify factors which can help in guiding changes and improving the system (Nickerson and Landauer, 1997, p. 16). Formative evaluation uses several of the techniques used in other types of evaluation, for instance surveys and interviews. Formative design evaluation can be contrasted to summative evaluation, which focuses on how good the outcome of something is. Summative evaluation has often been criticised for not resulting in valuable design guidelines.

You can’t rely upon descriptive data (Gould, 1988, p. 98), hence user testing is of great importance. User testing tries to get as close as possible to the actual use of the system. One method of achieving this is the use of demos and mock-ups in order to identify potential usability pitfalls before the system is implemented. On initial design, the average user interface has 40 flaws, of which half are easily identified simply by testing the system on two persons – a very illustrative example of the importance of user testing (Nickerson and Landauer, 1997, p. 17).

Our final approach is performance analysis. In the world of computerized systems, finding the right solution may be easy, whereas finding the right problem may be far more difficult. Performance analysis studies people doing information-processing tasks in order to find out what they do well and what they do poorly (Nickerson and Landauer, 1997, p. 17). The main
goal is to improve design with a focus on individual differences, time spent on the task-at-hand and potential error pitfalls. Methods include both measurements done in a lab-setting and real-world experiments.

Dix et al. discuss whether HCI has the properties of a craft in addition to a science (Dix et al., 2004, p. 6). There are several good arguments which support this view. Considering the wide nature of HCI both in methods, theories drawn from other fields and the fact that HCI touches upon development of almost any computer system, it becomes hard for one mind to grasp. Hence, reliance on experience and intuition may become important. As an example, consider the system designer being out in the field, studying people at their workplace. In addition to formal methods and specification, the designer’s social abilities (i.e. empathy, ability to interact with a wide array of personalities and ability to create an environment of trust in a very short time), could be of the utmost importance in achieving the desired result. In a second example, the design of user interfaces has its artistic properties. On a simple level, the composition of colors and pictures on a website, in addition to being guided by design specifications, presents the designer with an artistic challenge. On a more complicated level, the design of computer games can be a huge artistic undertaking, both in visual design and the creation of the world presented in the game.

I have identified six different approaches through which HCI can be studied. The borders between these approaches are far from fixed, and combining two or more of these could be a recipe for better results. Finally, there are arguments that can create a basis for claiming that HCI is a craft in addition to a science.

2.1.5 Why is HCI important?

Why is the field of HCI important and why should it be studied? The argument has been made that research on HCI is futile. Technology within the world of computers is moving so rapidly, research has little chance of affecting it (Nickerson and Landauer, 1997, p. 11-15). However, in this section I will present three arguments for why HCI is important after all.

First, research within HCI can reveal aspects of human tasks and activities which are in need of improvement, and help provide methods on how these improvements can be done. Poorly designed objects easily lead to frustration, lack of understanding and a higher probability of error (Norman, 1995, p. 5). A classical example is the omnipresent frustration of
understanding how the VCR actually works. Well designed objects are easy to interpret and understand. HCI can help reveal how objects can be improved and hence relieve the users of frustration, lack of understanding and help reduce error. Second, studying HCI can help us understand the effects technology and computers have on people’s productivity, job satisfaction, communication with other people and so forth.

Technology often has effects not predicted or intended by their inventor. For instance, it is highly unlikely that the creators of ARPANET (Internet’s predecessor) in the 1960s could foresee the quarrel between file sharers and the music industry found today, which their packet-switched network helped create the basis for. Technology has the potential to change life on our planet profoundly. We need to understand which potential effects technology has on society. Through HCI we could better understand the impact of computing on society and possibly steer the development towards humane outcomes (Baecker et al., 1995b, p. 903).

The third aspect concerns productivity. The effect of computer aids on worker efficiency is relatively slight. For instance, in a period from the 1970s to 1995, when the use of computers grew heavily, U.S. productivity grew at a little over 1 percent a year, compared to 2-3 percent annually in the previous century (Nickerson and Landauer, 1997, p. 12). Doing a bit better, word processing, which is the most widespread computer technology used in service industries, has been shown to produce a 50 percent increase in the amount of work done per hour (Nickerson and Landauer, 1997, p. 13). However, when mechanical equipment for spinning thread was introduced, it resulted in a 30,000 percent increase in efficiency in about the same amount of time computers have been around (Nickerson and Landauer, 1997, p. 13). In other words, the introduction of computers may not have been the efficiency-booster it could seem to be. This is where HCI might give a helping hand. Usefulness and usability is often where many basic problems lie. By making computerized systems more useful and usable, one could tighten the gap between the potential increase in efficiency offered by computerized systems and the actual lack of increase in efficiency present today.

There exist views which diminish the importance of HCI. However, by presenting the three arguments of improvement of design, effects of technology on society and the issue of productivity, I have tried to show that HCI should be considered an important contributor in the world of computerized systems.
2.2 Social Navigation

Social navigation does not have a single underlying theoretical framework (Höök, Benyon and Munro, 2003, p. 6). As with HCI in general, this does not imply that there do not exist any theoretical perspectives, but rather that social navigation contains a variety of theories and perspectives. In this section I will give the reader an introduction to the world of social navigation. I will start by defining the term social navigation, explain the physical-world metaphor and different context of usage. In the next chapters I will proceed to explain the differences between semantic, spatial and social navigation and then further investigate important aspects of social navigation theory, such as awareness and the distinction between direct and indirect social navigation. Finally, I will explain how social navigation is related to HCI.

I know nothing about cars. What would I do if I wanted to buy a car? A natural approach would be to contact somebody I know, someone who is familiar with cars, and ask them. Have you ever been on holiday and in a position where you had to choose between the half-empty and the well-populated restaurant and then ended up choosing the populated one based on the thought that most people would most likely go to the best one? These are all examples of physical-world social navigation. According to what most social navigation literature regard as the original definition of the term, social navigation is moving “towards” a cluster of other people, or selecting objects because others have been examining them (Dourish and Chalmers, 1994, p1). I chose a car based on my friend’s recommendation, you chose the restaurant based on how many people were sitting inside of it. Social navigation is hence based on the idea that information about other people or about other people’s activities can be of great value to other individuals in the conduct of performing an activity (Dourish, 2003, p. 278).

Social navigation in the digital world is inspired by how people gather information in the physical world, as illustrated by the initial examples in this chapter. Wexelblat and Maes point out how information in the digital world comes with no history (Wexelblat and Maes, 1999, p. 270). In the physical world, we make extensive use of traces, such as asking a friend, on which we base our choices or to find matters of importance and interest. We make use of social navigation every day in a multitude of ways, without conscious reflection. Work within
the field of social navigation takes the physical-world metaphor in varying degrees. On the one side, one can choose one single aspect of physical-world social navigation and try to adapt this to a single, useful feature, such as the rating-of-content features flourishing on the Web today. In these cases a simple physical-world action (asking someone you know what they thought about a specific item, such as a movie) is replicated on the web. As opposed to the physical world, the system can be used by a great number of individuals, there can be several thousand individuals expressing their opinion about a single item (not just your friend). A different approach is found in the experiments of McGrath and Munro 2003, in which they try to catch the physical-world metaphor more literally (McGrath and Munro 2003). In order to capture the informal aspects of cooperative work, such as work outside formal routines and seemingly purposeless chatter between co-workers or between seller and potential customer (such chatter can be a good way of establishing contacts, and hence not purposeless after all), they attempt to replicate this physical-world phenomenon in a computerized 3D environment. Both examples above illustrate how social navigation techniques are implemented based on the physical-world metaphor, how varying aspects of the metaphor are used and the degree of literacy.

Let’s take a closer look at what social navigation is not. Walking a path in the forest, created by several other individuals walking there before you, is an example of physical-world social navigation. Driving your car on the highway is not social navigation. It was the intention of the constructor for you to drive on the highway. Contrary to this, the path in the forest was not created with somebody’s underlying intention of other people to use it. Similarly, when you read product information on the Internet from the website of the company which is behind the product, it is the company’s intention to supply this information in order to inform you about the product (and hopefully make you buy it), and hence this activity can not be considered as social navigation. On the other hand, if you read reviews of the same product written by users of a web forum, the intention of the creators of the forum is not to make you buy a specific product, but rather to supply users the service of being able to write any product review. The latter is social navigation, you choose to act (buy or not buy) based on what other people have experienced before you.

An important aspect of social navigation on the Web is that through implementing social navigation techniques, the creator of a service loses some control by opening the possibility of users creating their own content or exercising control over existing content. When a web
service is built, the designer usually has some sort of intention behind the creation of the system. However, through opening social navigation services and user-created content, the outcome is not necessarily what the designer intended (Dieberger et al., 2000, p. 39). An example from 2007 illustrates this point. In May 2007, several users of Digg (www.digg.com), a website which enables users to submit and vote on news, articles, and the like, several users posted links to websites revealing the key for the HD-DVD format. Digg responded by trying to remove these links (since publishing the encryption key is illegal), whereupon the Digg community of users revolted by massively posting new links and voting on items containing references to the link. The result was that the site's entire homepage was covered with links to the HD-DVD code (Greenberg, 2007). Digg finally gave up attempts to remove references to the encryption key, stating that we hear you, and effective immediately we won’t delete stories or comments containing the code and will deal with whatever the consequences might be (Rose, 2007). This type of usage was probably not what the designers of Digg had in mind when it was created.

Social navigation techniques can be implemented in a wide range of different environments (Höök, Benyon and Munro, 2003, p. 5). The most common techniques are textual implementations, such as those used by most websites. Online chat is also a good example of textual social navigation. It is not uncommon to combine text with icons (often called “emoticons”), in order to enable the users to express emotions language cannot convey. Sound and video are two other examples. Social navigation can take place both through sound, such as a podcast, and video, for example the various recordings of other people’s bad experiences with customer support found on a wide array of websites. Obviously, phone and video calls through the Internet also constitute a channel for social navigation, but this isn’t really revolutionary, considering the phone’s been around for quite some time. Last, you have 3D animated environments, such as the multitude of online games which requires cooperation between the players.

The most common definition of the term social navigation is doing something because others have done it before you. Social navigation is heavily influenced by the physical-world metaphor, but this metaphor is used in varying degrees. Implementing social navigation

---

2 The HD-DVD encryption key can be used to decrypt the discs, and hence be used to copy their content.
techniques can be done in a wide array of different environments, ranging from textual environments to three-dimensional worlds.

2.2.1 Semantic, Spatial and Social Navigation
When navigating the Internet and the Web, people use several different techniques. This section presents three of them, semantic navigation, spatial navigation and social navigation.

A very common method of navigation on the Web is by grouping objects (i.e. web pages) according to similarity. This technique is called semantic navigation. Semantic navigation utilizes the underlying similarities between information objects and renders a relationship between them (Dourish and Chalmers, 1994, p.1). A straight-forward example of this is a simple search engine. By entering a word, say “car”, the search engine traverses its database, retrieves a selection of web pages relating to the word “car” and displays the result. Grouping web pages according to a semantic relationship presents the user with a simple and straightforward way of navigation. The user can relate to the semantic relationships found between information objects, hence making navigation logical and simple. For example, a user would probably not be very surprised if the search result for the word “car” displayed a link to the site of BMW. There is a clear semantic relationship between cars and BMW. On the other hand, semantic relationships can be troublesome. I will present three problem areas related to semantic navigation. First, an obvious problem arises when a word has more than one meaning. What happens, for instance, if you want to learn more about pool? A semantic search engine would not know whether you searched for the game pool or swimming pools. Wittgenstein states that “the meaning of a word is its use in the language” (cited in Chalmers, 2003, p. 321). Due to the versatility of languages, structuring information according to content of a document will have shortcomings. Coyle et al. points at a second challenge, namely the users’ inability of effectively expressing their needs in a query (Coyle et al., 2007, p. 54). As an illustrative example, take the simple act of spelling a word wrong. A semantic system would try to find relationships between the misspelled word and other documents. In other words, there is an input problem. A semantic system takes the input quite literally, it does not capture the users’ actual intention behind a query. If a semantic navigation environment is given the wrong input, it will return the wrong output. Last, we have the issue of algorithms. Wegner writes:
Algorithms are metaphorically dumb and blind because they cannot adapt interactively while they compute. They are autistic in performing tasks according to rules rather than through interaction. In contrast, interactive systems are grounded in an external reality both more demanding and richer in behavior than the rulebased world of noninteractive algorithms.

(Wegner, 1997, p. 82)

In other words, according to Wegner, algorithms are simply not capable of capturing the complexity of human interactions. They are “blind”, do exactly what they are told, without regard to a changing environment. Human interaction cannot be reduced to the point where it can be understood by an algorithm.

A second method of navigation is spatial navigation. As the term implies, spatial navigation is navigation by movement in space. An example of a service which uses spatial navigation is IMVU (www.imvu.com). This service utilizes spatial navigation by implementing Internet-based chat in a 3D environment. The most extensive usage of spatial navigation however, is found in the entertainment industry, for instance massive multiplayer online role-playing games (MMORPG), such as World of Warcraft (www.worldofwarcraft.com). Spatial navigation has some obvious shortcomings as a tool for navigation. It would be rather cumbersome to run around in a three-dimensional “Google-land”, trying to find the classifieds of the local paper. It does however have advantages, such as being able to see other users, facial expressions and gestures, which cannot easily be expressed using semantic navigation.

Social navigation, as opposed to the above techniques, creates navigation based on what other users have done before you. Social navigation is not a competitor to the above techniques, nor a solution to all their shortcomings, but rather a complementary navigation technique. It does however offer opportunities for navigation which semantic and spatial navigation does not. For instance, semantic navigation does not give users the possibility of judging an item based on other users’ experiences. On the other hand, semantic and spatial navigation have properties and advantages social navigation does not have. For instance, in an experiment done by Wexelblat, in which they try to make a general navigational system for the Web using social navigation techniques such as heavy reliance on metaphors, they find social navigation not to have a significant effect on browsing (Wexelblat, 2003).
I want now to turn back to Wegner’s above quotation on algorithms. In the last sentence he states that *interactive systems are grounded in an external reality*. Social navigation techniques can be regarded as what Wegner calls “interactive systems”. Social navigation environments are populated by several users. You are not alone. The users interact, and hence social navigation environments constitute what Wegner calls “interactive systems”. He further elaborates on how interactive systems are grounded in an external reality, and in that being richer in behaviour as compared to ruled-based (algorithmic) systems. Due to the multitude of users in a social navigation environment, and their ability to affect their surroundings, these environments are grounded in an external reality (the users and their actions) and hence enable the “richer behaviour” Wegner is longing for. To sum up, social navigation techniques can be very helpful to solve the challenge of semantic navigation’s autistic non-interactive algorithms.

This section has so far described the three techniques of navigation as separate. The purpose is to shed light on the distinction between the techniques. But, the three techniques are very often used simultaneously. For instance, IMVU allows its users to see each other in a 3D environment. However, and this is also the reason why 3D environments often constitute a grey-zone between spatial and social navigation, this service enables users to move towards a crowd of other users, which is social navigation in practice. Many websites which rely heavily on social navigation techniques also offer a search option, hence creating a combination of social and semantic navigation.

I have presented three common types of navigation on the Web, semantic, spatial and social navigation. All of the three systems of navigation presented in this chapter have advantages and disadvantages. Combining them is common, and can be a very useful approach

### 2.2.2 Space and Place

The web is a “lonely” place. If you read a newspaper article online, you are viewing the site “alone”, as if you were reading the article on paper. However, it is very likely that several other users right now are viewing the site you are currently visiting. This section will present how social navigation can help bridge the gap between “loneliness” and awareness of other users.
Social navigation techniques have the ability to change a *space* into a *place*. A place is defined as a space with meaning (Maglio, Barrett and Farrell, 2003, p. 249). For instance, in the physical world a church is a place. It has *meaning* in the sense that there are certain activities and actions naturally related to the place, i.e. worshipping. Similarly, a pub is a place, whose meaning could include the activity of drinking beer. An example of a *space* could be the newspaper site described above. It has no meaning related to it; the site consists only of text on a screen. Through the use of built-in social navigation techniques on the web, the site can be transformed from a space into a place. Social navigation creates *interactions between users* and makes the users *aware of each other* (Maglio, Barrett and Farrell, 2003, p. 250). As an example, our newspaper site can be turned into a place through for instance a discussion board. This discussion board facilitates the exchange of opinions between users, hence creating a place of awareness between users.

The notion of turning a Web space into a place creates a new set of interesting perspectives. Dourish points out how some populated places automatically imply a shared understanding of appropriateness (Dourish, 2003, p. 284-288). Similarly, Dieberger shows how people associate social connotations with space (Dieberger, 2003, p. 294). In the physical world, this is old news (you should take your hat off when you enter a church). An example from the digital world is found in the MMORPG game World of Warcraft, where asking other people for money (begging) is frowned upon. It seems, due to the transformation of a digital realm from space to place, physical-world social connotations and rules of behaviour are introduced into the digital world.

Turning spaces into places through the use of social navigation techniques can have several beneficial aspects. Dieberger points out how social navigation gives the user the option of contributing in a place, and hence a sense of ownership (Dieberger et al., 2000, p. 43-44). This again could make the user stay longer on the site and visit more frequently. In addition, though moving a bit away from the sole notion of place and into social navigation in general, these techniques can help users find more relevant information and more quality information as compared to semantic navigation.

Turning space into a place is an important aspect of social navigation, and presents a whole new perspective on navigation compared to semantic and spatial navigation. A place can create awareness of other users, social connotations and a sense of ownership.
2.2.3 Direct, Indirect and Social Texture

We can make a distinction between *direct* and *indirect* social navigation (Höök, Benyon and Munro, 2003, p. 5). Direct social navigation is when there is a real-time contact between the actors, such as the built-in chat found in many MMORPG. Indirect social navigation is when there is no direct contact between the actors, such as a written book review. These two examples are both pretty straight-forward, but *social texture* requires a bit more explaining.

Say you’re in a library browsing the shelves and stumble upon a worn book. Its edges are torn, some of the pages are about to fall out and there are notes and scribbling from previous lenders. These are examples of social texture, cues left by other people. By the wear and tear, you can conclude that the book is popular and might be worth reading. A good example of digital-world social texture is tags, which is the ability for users to add a few words describing an item which again can be used by other individuals to navigate the information space. Social texture is not one hundred percent distinct from direct and indirect social navigation (one can say that tags are indirect social navigation), but is a concept not caught by the distinction between direct and indirect social navigation, and hence deserves its own review.

We can make a distinction between direct social navigation, indirect social navigation and social texture. Direct and indirect social navigation are clearly distinguishable whereas the lines between social texture and direct/indirect social navigation are blurred.

2.2.4 Social Navigation and HCI

Social navigation is regarded as a subfield of HCI. How is social navigation associated with HCI? This section seeks to link the two fields together and explain how social navigation techniques can be a very valuable tool.

As noted in the chapter on HCI, one of the major challenges of computerized systems today is a lack of effectiveness. By focusing on usefulness and usability in the design of computer systems, we might help solve this problem. The below quote from Nickerson and Landauer illustrates this challenge.
The challenge facing HCI is to be able to use knowledge about actual effectiveness in helping people achieve their goals to make new applications not only novel, exciting and technically impressive, but useful, usable and socially valuable as well.


Social navigation techniques can help achieve the goals of usefulness, usability and social value. Using the definition by Dourish and Chalmers, social navigation is per se user centred. Through choosing objects because others have been looking at them before, usefulness and usability of a computerized system can be increased. You do not have to make the same mistakes others did before you, you can base your choice on a pool of existing experiences. To exemplify, let’s take a brief look at the website The Internet Movie Database (www.imdb.com). This site enables its users to vote on (almost) any movie, and displays an average rating based on these votes. If you are considering renting a specific movie, this rating instantly gives you an idea of how other individuals liked this movie. Thus, social navigation helps increase both usefulness and usability of this site. We have already discussed how social navigation helps transform a space into a place, and how this enables users to contribute and gain a sense of ownership. This aspect of social navigation runs straight into the arms of Nickerson and Landauer when they state in the quote above that HCI should focus on social value. By implementing social navigation techniques, a computerized system can help create a community, in which the users can feel a sense of belonging, thus giving the system social value. Achieving effectiveness is by some considered a major challenge within HCI. Social navigation techniques can help achieve this goal through usefulness, usability and social value.

2.3 Social Navigation Rating Techniques
This section will present the subject of social navigation rating techniques. The subject will be presented in two sub-sections. I will first take a look at the subject of collaborative filtering (CF), also known as recommender systems and second, take a closer look at user-written reviews. Some readers might argue that written reviews are not a rating technique. The reason this issue is included here is that written reviews will be included in my research, and hence deserve a thorough explanation. On the one hand, written reviews are not a form of rating if you define rating as putting something on a scale. On the other hand, one could claim that
when your write a review, i.e. how much you disliked the latest James Bond movie, there is an inherit rating (“disliked”) in the review. There are differences, there are similarities, and that’s why I’ve included them in the same part, but split in two chapters.

2.3.1 Collaborative Filtering

*Launch.com allows users to rate songs on a scale from 0 to 100, plus a control we call the “Britney Spears button” that allows the user to never hear a particular song again.*

(Cosley et al., 2003, p. 586)

As the amount of information on the Internet, or even on a single web page, increases, users are faced with the problem of information overload. There is simply too much information out there, and finding the information one’s looking for becomes exceedingly complicated. Collaborative filtering (CF) is a set of techniques aimed at directing users towards the information they find most valuable (Konstan and Riedl, 2003, p. 44). The users of a computerized service are given the option of in some way rating an item and the CF system then uses these ratings to recommend items to other users. An item could be more or less anything, ranging from songs, videos and consumer electronics to travel destinations and even other users.

Collaborative filtering can be closely associated with the physical-world metaphor. In everyday life, when faced with the problem of making a choice without enough personal experience, we often rely on recommendations from other people, word of mouth, movie and book reviews in newspapers and similar channels of information. Recommender systems can assist and augment this physical-world social process (Resnick and Varian, 1997, p. 56). Hence, collaborative filtering is a set of techniques well within the field of social navigation.

To further illustrate collaborative filtering, I will present three very common approaches to the implementation of these techniques. The below selection of implementations are prime examples of the usefulness of social navigation rating techniques as a navigational tool. An exceedingly common area of usage is in entertainment sites, where users can rate the content.
A good example is Break (www.break.com), which is a site where users can view more or less entertaining video clips. Any user can rate any video on a scale from 1 to 5. For each video, the current average rating is shown as a row of stars, hence giving other users an idea of how well previous viewers liked this specific video. In addition, there is a “Top Rated” menu, which is dynamically generated based on which video clips have the highest average ratings. The picture below shows how an item’s rating is presented on this specific website. In the lower left corner, the average rating is presented as a row of stars. The number to the right of the stars represents the total number of votes.

![Picture 1: Rating an item at www.break.com.](image)

In our second example we’ll take a look at an almost famous example of a collaborative filtering system, the one employed by Amazon (www.amazon.com). Based on yours and other users’ history of book purchases and other users’ ratings on the same 1 to 5 scale as Break uses, this CF system compiles a targeted list of book recommendations for each user and/or suggests other books based on the book you’re currently viewing. The difference between Amazon and Break is that the CF system implemented by Break presents an average common view of the item rated, whereas Amazon gives targeted, individual recommendations for each user. The picture below illustrates how Amazon recommends books (bottom row in the picture) based on which book you are currently viewing.
The last example is from the Norwegian auction site QXL (www.qxl.no). This is a site where users buy and sell products and services (very akin to eBay). Whenever a transaction between a seller and bidder is complete, they are each given the option of rating each other, either “good”, “neutral” or “bad”. This example differs from the two above, in that the items rated are humans, and that the main purpose behind the system is to build trust. Based on other users’ experiences with a specific seller, a potential bidder is able to make a judgement on how trustworthy this person is – trusting a seller because others have shown before that the seller is reliable. The picture below illustrates how rating is done. The diamonds shows that this user has performed a large number of sales. The number 432 shows the total number of ratings this user has. The value 99.31% is the average number of good ratings, and the row of stars shows how many transactions the user has done in total.
In general, collaborative filtering systems can be divided into three parts, resembling the classical input – function – output structure found in programming.

1. **Input**
   (i.e. the rating of an item on a scale from 1 to 5).

2. **Algorithm**
   (i.e. calculation of an item’s average score).

3. **Output**
   Recommendation.

**Picture 4: Structure of a CF system.**

1. The system is given input, i.e. users rate an item on a scale or a user’s purchase history is recorded.
2. An algorithm calculates a recommendation, i.e. an average rating or, through comparing input from other users, making a targeted recommendation.
3. Displaying the recommendation created by the algorithm.

Collaborative filtering systems come in all sorts of shapes and sizes. In the remainder of this chapter, I will describe various properties of CF systems. Whenever CF systems are mentioned in social navigation literature, it is mainly with reference to a system akin to that of Amazon described above, where users receive targeted recommendations. However, there is a difference between what I’ve coined personal and impersonal recommendations. A personal CF system is when each user receives a targeted recommendation aimed at a single individual user, as with the Amazon system described above. An impersonal CF system is when each user is presented a common average view of an item, such as with the Break example above. Technically, the differences between these systems are vast. Making an impersonal recommendation system requires a course in basic programming, and the ability to calculate averages. When it comes to personal CF systems, however, the algorithm behind the recommendations can be extraordinarily complicated. Explaining the inner workings of these algorithms requires a thesis (or several?) in itself, but the interested reader could take a look at “Evaluating Collaborative Filtering Recommender Systems” (Herlocker et al., 2004). Within the category of personal recommendations, there is the distinction between user-based and item-based algorithms (Chirita, Nejdl and Zamfir, 2005, p. 67). User-based algorithms build for each user a neighbourhood of other users with similar opinions. Ratings from these users are then employed to generate recommendations for the target user. Item-based algorithms, on
the other hand, are based on similarities between the items themselves and use these similarities to calculate recommendations.

A second aspect of CF systems concerns how an item is rated. There is a distinction between active and passive data recording (Wexelblat and Maes, 1999, p. 271). This distinction is also known under the terms explicit rating (active) and implicit rating (passive) (Konstan and Riedl, 2003, p. 58). Active data recording is when the user has to make some effort in order to place a vote, such as with the rating of videos on the Break website. Passive data recording is when a CF system records information without the user having to do anything, such as how Amazon records data based on a user’s purchasing history. The advantage of passive data recording, and hence disadvantage of active data recording, is naturally that the system doesn’t have to interfere with the users’ actions. A disadvantage of passive data recording is illustrated by the fact that users of a web shop doesn’t necessarily buy items they like (Konstan and Riedl, 2003, p. 58). Say for example you’re buying some books from an online book store as a gift for your mother, and this book store has a CF system which creates recommendations based on a calculated perception of your taste, which again relies on your purchase history. Problem is that “your” taste is now a combination of your and your mother’s taste, hence making the CF system’s recommendations unreliable.

A variation of passive rating is to record how long an item was viewed in order to determine how well the user actually liked the item (Coyle et al., 2007, p. 53). Take for example a site which records viewers of video clips. A possible approach is to determine how popular an item is by how many users clicked on the link. A problem arises when a user finds the video clip boring and browses away from the page after just a few seconds. In this case, the measurement of popularity becomes unreliable. A better approach would be to record how long the user actually spent watching, in order to determine whether (s)he actually liked the video clip or not. This method was used by Colye et al. in order to determine how relevant a set of articles was depending on how long students used on viewing the article (Coyle et al., 2007, p. 53). Combined with several other social navigation and non-social navigation tools used in their project, it helped the users better understand the relevance of annotated information from the perspective of their own community’s interests (Coyle et al., 2007, p. 60).
A third aspect of CF systems is the idea of them being self-correcting. This is closely associated with what Coyle et al. calls community wisdom (Coyle et al. 2007). The idea of self-correcting CF systems is best illustrated by an example. If an item with an artificially high rating was inserted into a CF environment, the “community wisdom” of the users would result in the item (over time) being reduced to its appropriate rating. In this sense, the CF system can be said to be self-correcting. The idea of a CF system being self-correcting gives it great strength. Through the filtering done by masses, an item’s “true” rating will emerge, giving the users a base of trust in the recommendation. This of course relies on the community actually being wise, but more on that in the chapter on challenges of social navigation rating techniques.

Collaborative filtering is a method for directing user towards information they find most valuable. I have identified three different areas of usage, simple average rating of an item, targeted recommendations and the case of rating other users. We can distinguish between personal and impersonal recommendations, and personal recommendations can again be divided into item-based and user-based. When collecting data from users there are two approaches, namely active and passive data collecting. An important aspect within the theory surrounding CF systems is the idea that they are self-correcting.

2.3.2 Written Reviews
A fourth aspect of CF systems is degree of intimacy. The information gathered about an item can be more or less social or more or less personal (Wexelblat and Maes, 1999, p. 271). Social information describes “what has been done?”. For instance, a book rated on average four of five stars by twelve hundred users is an example of this. Personal information, on the other hand, is characterized by “what did I do?”. An example of this is written reviews.

Written reviews have been around since long before the Internet existed, such as for instance, movie and book reviews in newspapers and magazines. As opposed to written reviews in newspapers and magazines, usually authored by individuals with some sort of expertise within the field in question, the Internet has created the possibility of anyone writing reviews about pretty much everything.
Written reviews are often found in combination with rating-on-a-scale systems, and will hence be a topic in this work. For instance, Amazon offers its users the opportunity of writing reviews in addition to rating an item on a scale. Combining written reviews and ratings on a scale can have a complementary effect. A rating on a scale give a user a simple and quick way of judging other users’ perception of an item, but does not give a deeper explanation of an item’s content, i.e. the plot of a movie (Burke, 2002, p. 336). For instance, can such a complex thing as a book really be represented by a numeric value between 1 and 5? Written reviews give the user the possibility of gaining a deeper understanding into other users’ perception of the item. Swearingen and Sinha shows that users like to have more information about an item, such as reviews by other users in order to increase perceived usefulness and ease of use of the recommender system (Swearingen and Sinha, 2002). On the downside, reading reviews requires more effort from the user as compared to a glance at the average rating.

One can distinguish between two types of written reviews, expert reviews and reviews written by just anyone. Expert reviews are reviews written by an individual who has some sort of expertise within the field, such as the newspaper example above. Expert reviews can induce a higher degree of trust in the reader, since the author is known to have expertise within the field in question. The limits of expert reviews are that there are a limited number of experts and those experts may charge for writing their reviews. The second type of reviews is those which are written by just anyone, or in other words, not a (known) expert within the field. These have the disadvantage of possibly being less trustworthy, but the advantage of being usually plentiful and free-of-charge.

Rating items on a scale does not necessarily give other users a proper perception of an item. Written reviews can augment rating systems through giving the user a deeper understanding of the item in question. One can distinguish between two types of written reviews, expert reviews and reviews written by any user.

2.4 Challenges of Social Navigation Rating Techniques
This section will present eight different challenges of collaborative filtering. The challenges presented next are a compilation of known issues found in social navigation theory and problem areas elaborated by other researchers. All chapters will have the following structure:
First comes a presentation of the problem area. After this I will present possible solutions to the problem described. At last, I will describe why or why not this is an issue which will be further elaborated in this work. Challenges within the following eight challenges will be presented:

1. Algorithm, the technical core of a CF system.
2. Shilling attacks, faking ratings.
3. Barrier to entry, how new items gain ratings.
4. Conformity, how the existing rating can affect a new rating.
5. Rating scales, how the rating scale can affect ratings.
6. Confidence, building trust in the recommendation.
7. Fading, how ratings age.
8. Privacy, protecting the privacy of the individual user.

2.4.1 Algorithm
One could say that the algorithm is the core of a CF system. Based on the input of one or more users, the algorithm outputs recommendations to the same or to other users. As presented above, there exists a variety of different CF algorithms, ranging from simple calculation of average rating, on which you can present a common average validation of an item, to complicated statistical analysis such as Bayesian algorithms on which you can present personal recommendations targeted at a single user.

A basic problem related to CF algorithms is how well the predictions they produce match the actual taste of the user, or in other words, the accuracy of the algorithm. (This does not directly apply to simple algorithms that add up scores and present an average, as the goal is to present a common average validation, not to produce individually targeted recommendations.) Most of the research within CF focuses on the issue of accuracy (Cosley et al., 2003; Herlocker et al., 2004; Swearingen and Sinha, 2002; O'Mahony et al., 2004 and many more), to the extent where most of the accuracy problem has been solved (Cosley et al., 2003, p. 592). The study of CF algorithms is a field of its own, and outside the scope of this thesis.

However, we’re not going to let the algorithm go that easily. A second problem is the issue of the correctness of input data. It is basic computer knowledge that if an algorithm is fed bogus
data, it will output bogus data. Generally, it does not matter how well the CF algorithms are
designed, if the inputted data remains faulted. To some extent, bogus input data can be
discovered by special algorithms, which will be one of the topics of the following chapter.

**2.4.2 Shilling Attacks**

A shilling attack is when a user or users deliberately manipulate a CF system in order to credit
or discredit an item. Shilling attacks are also known under the name *spoofing* (Dieberger et
al., 2000, p. 42). A simple example could be if a company gives its own products high ratings
and false user reviews on a product recommendation site in order to gain an advantage over
its competitors.

There are several physical- and digital-world examples of manipulation akin to shilling
attacks. Sony admitted in June 2001 to using fake quotes to promote its films (BBC, 2001). In
September 2003, eBay made changes to its feedback system, after experiencing trouble with
users buying or trading feedback from other users (AuctionBytes, 2003). In March 2002,
Amazon deactivated one of its advising features due to “commercial abuse” (Wired, 2002),
and in a second Amazon incident, due to what was described as a “prank”, the Amazon
recommender system was tricked into recommending a rather inappropriate sex manual to
readers of a Christian spiritual guide, the latter, according to security consultant Richard
Smith (quoted in the article), *made him* also wonder how much the Amazon
recommendation system is being hacked by authors and publishers as a new marketing tool
(News.com, 2002). These examples illustrate a real problem which absolutely requires the
attention of both designers and researchers.

One can distinguish between two types of shilling attacks, *push attack* and *nuke attack*
(Chirita, Nejdl and Zamfir, 2005, p. 67). In a push attack, the purpose is to artificially increase
the rating of an item. A nuke attack is the opposite, an attempt to artificially decrease the
rating of an item. An example could be a company writing bad reviews on its competitors’
products.

Different CF systems have varying degree of vulnerability towards shilling attacks. Naturally,
items with few (or no) ratings are more exposed. The fewer ratings an item has, the easier it is
to perform a push/nuke attack on that item, simply because fewer, or possibly even just one,
new rating greatly affects the product’s existing (or non-existent) rating. If an item is already rated by thousands of other users, it’s hard to have a significant influence on this item’s rating. This argument is recognised in the research of Chirita et al. in which paper they state *our experiments have shown that many items in the system are rated by only a few users, and their ratings can be pushed very easily* (Chirita, Nejdl and Zamfir, 2005, p. 67). It is worth noting that the combination of shilling attacks and lack of ratings are a threat to the idea of CF systems being self correcting. The combination may actually produce the exact opposite; self falsifying CF systems. If there are few (or no) ratings, and an actor with the incentive to insert fake ratings, the result could be a (partially) falsified system of “recommendations”.

Protecting against shilling attacks can be extraordinarily difficult. Take for example the following (simple, but illustrative) scenario: A worker in a company finds their new product on a product recommendation website. The product is not yet rated. The diligent worker inserts a high rating and writes a good review, with the sole purpose of creating a positive image of their product. How can one possibly discover that this rating is false? For all we know, this rating is absolutely legitimate, written by a satisfied customer\(^3\).

Luckily, there are techniques which can help protect against shilling attacks. First, this challenge can be partially prevented at the technical level. For Heaven’s sake make sure a single user cannot rate the same item more than once, in order to prevent a user from clicking like crazy on a desired item, or worse, a bot attack. In addition, track all users’ ratings so, in case of a suspected shilling attack, all ratings from a single user can be easily removed. Second, in their research on shilling attacks, Lam and Riedl show that new items are especially vulnerable (Lam and Riedl, 2004, p. 401). As a solution to this problem, they suggest that new items could be rated from a trusted source. For instance, the ratings for a new item could be seeded with ratings from professional critics or from trusted volunteers. Third, item based algorithms appear to be more resistant to shilling attacks (Lam and Riedl, 2004, p. 401). Hence item-item algorithms are preferable to user-user algorithms, and should be the choice of implementation if possible. A fourth technique is being on the lookout for abrupt changes in the rating of a single item. I have found two different papers presenting a somewhat conflicting view on the efficiency of such algorithms. Interestingly, both papers use

\(^3\) For the technically minded: The worker did it from home, so there’s no way an IP address can be traced back to the workplace.
the same case in their research, the MovieLens online movie database, and both used algorithms to try to detect patterns of a shilling attack. The first paper reports mixed results, concluding that algorithms have the potential of discovering a shilling attack, but that effective shilling attacks cannot be detected that easily (Lam and Riedl, 2004, p. 401). According to the results of the second paper, attackers exhibit special noticeable rating patterns which can be effectively detected by the algorithm created by the authors (Chirita, Nejdl and Zamfir, 2005, p. 73). Due to the conflicting results, concluding on how efficient such algorithms are is difficult. We can however with some certainty say that algorithms which detect patterns of shilling attacks can be useful and have the potential to discover when a system is under attack.

The challenge of shilling attacks is not an easy one. In my research, the focus will be on a system’s vulnerability to these kinds of attacks, evaluating the degree to which a CF system can be intentionally manipulated.

2.4.3 Barrier to Entry
The barrier to entry problem occurs whenever there is an item which is not rated. For example, if a new item is inserted in a CF environment, and this item’s rating is set to “none”, it easily drowns in a sea of already rated items. O’Mahony et al. ironically points out that for an item that is not rated, the recommendation “recommended to no one” and “recommended to everyone” may both be correct, yet contradictory (O’Mahony et al., 2004, p. 345). The barrier to entry problem is fuelled by the inherit mechanisms in CF techniques, namely that a highly rated item is more likely to be exposed to a higher degree in the CF environment as opposed to a low-rated item. In their research on an online recipe recommendation system, Svensson et al. shows this effect in action in describing how people tended to move towards a recipe collection simply because there were many other people there (Svensson et al., 2005). Interestingly, the basic idea behind social navigation is exactly this, doing something because others have done it before you, and hence the barrier-to-entry challenge is inherent within social navigation systems themselves.

One can roughly divide the barrier to entry problem into three different categories. The first category is when a web site is new and no items are rated. Depending on how heavily the site relies on collaborative filtering techniques, boot-strapping the CF environment can be a major
problem. For instance, on web sites which build dynamically generated menus based on user input, such as “Most Viewed” or “Top Rated”, the barrier to entry could pose a major problem as the site goes on-line for the first time. Other sites, such as web shops which offer a product rating option, the problem is less severe. The web site is still functional without the user ratings.

The second category is when new items are inserted in a CF environment and are not yet rated. Since no rating often is equivalent to bottom of the list, the newly added item will struggle to achieve a good rating even though it might very well be one of the most popular items ever added on the site. A special variation of this problem exists in web-based reputation management systems, that is, CF environments where users themselves are rated. An example of such a system is the online auction system QXL mentioned earlier. New users of such systems often start without any rating at all, and hence do not have any reputation, neither good nor bad. Several studies have looked into this challenge. Landon and Smith show that reputation has a large impact on the willingness of consumers to pay (Landon and Smith, 1997, p. 289). A study done by Malaga on web-based reputation management systems shows that many users of the eBay online auction system (www.ebay.com) will not deal with individuals with a low reputation score (Malaga, 2001, p. 407). A separate study done by Houser and Wooders (also on eBay) further strengthens the reputation problem by showing that seller reputation (but not bidder reputation) is a statistically and economically significant determinant of auction prices (Houser and Wooders, 2000, p. 367).

The third category relates to the simple lack of a sufficient number of ratings, that is, a large part of items on a web site is not rated. As an example, let’s make some trouble for the seemingly-blissful web shop example above. Let’s say a situation occurs in which only half of the products are actually rated. In this scenario, user recommendations become a lot less valuable, since the ratings do not encompass the entire product spectrum. In addition, dynamically generated user recommendation menus become a lot less trustworthy again due to the fact that they are not based on a review of the entire product spectrum of the web shop. To sum up, the problem of barrier to entry may pose a great challenge to CF systems.

There are several different possible solutions to the problems described above. One reason why items lack ratings is the users’ lack of incentive to rate, also known as the free-rider problem (Dieberger, 2003, p. 300). This challenge can be solved by offering the user some
kind of reward when the user rates an item. However, awarding must be done with caution, in order to avoid users senselessly rating items in order to receive the award offered. A different approach is to provide a small increase in reputation whenever a user provides reputation feedback to others (Malaga, 2001, p. 412). The usage of this solution is however limited and will most likely remain useful only in reputation management systems since it can only be applied in environments where there exists an option of rating the users themselves.

Another solution is found in the differentiation between active and passive data recording (Wexelblat and Maes, 1999, p. 271). Active data recording implies that the user explicitly rates an item, i.e. by rating the item on a scale. Passive data recording relieves the user of explicitly rating items, and hence could increase the number of ratings. On the other hand, not all items can be rated passively, for example how satisfied a customer was with the product that was bought. Further, passive data recording faces a problem when it comes to determining a single user’s taste, as exemplified by the gift-to-your-mother story in the theory chapter.

A third solution, targeted at new items inserted in a CF environment where other items are already rated, is proposed by Malaga in his study of reputation management systems (Malaga, 2001, p. 411). This solution is proposed for rating other users, but could be extended into rating other items. The idea is simple: When a new item is inserted into a CF environment, it receives a default rating. This default rating should not be a fixed value, since users easily may recognise the fixed value of new items, but rather the mean rating of all existing items of the same type. The newly inserted item could then have an initial fair chance in the competition between the other items within the CF environment.

In his quest for solving the dilemma of newly added items in CF environments, Malaga proposes a second solution. This is simply not to show an item’s score until it has received a certain number of ratings (Malaga, 2001, p. 411). A somewhat similar idea, but conceived from a psychological viewpoint, is presented by Cosley et al. when they suggest the option of hiding existing ratings when the user rates an item (Cosley et al., 2003, p. 591). The idea is the same, but the justifications differ. Malaga’s basis for hiding the rating is to make it easier for new items to achieve a decent score within a CF environment, whereas Cosley’s justification is to avoid users being influenced by an existing rating when themselves rating the same item (more on Cosley later in this chapter).
Finally, there’s a solution which I surprisingly have not found in any literature, even though it’s been around the Internet for years, namely the “New”-list. The idea is simple. Whenever a new item is inserted into a CF environment, it appears on a list of newly inserted items. This list is made easily accessible in the site’s menu structure (or even is the first page of the site), and hence, through this increased exposure of the infant items, the path towards the world of rated items should be easier to walk.

The barrier to entry problem poses questions I will look more closely at in my research, focusing on newly inserted items, if there is a general lack of ratings and which measures are taken to facilitate users’ rating.

2.4.4 Conformity
In the next area of problems, we move into the world of psychology. In their article, “Is Seeing Believing: How Recommender Interfaces Affect Users’ Opinions”, Cosley et al. states the following:

The psychological literature on conformity suggests that in the course of helping people make choices, these systems probably affect users’ opinions of the items. If opinions are influenced by recommendations, they might be less valuable for making recommendations for other users (Cosley et al., 2003, p. 585).

This is potentially a grave threat to CF systems. The paramount idea behind collaborative filtering is to filter out unwanted items so the preferred items remain in a digital world of information overload. CF systems are believed to be self correcting, i.e. if an artificially high rating was given to an item, other users would give true ratings for that item and cause it not to be recommended any more. If opinions are influenced by the already existing rating, the idea of self-correction loses validity, and hence the entire idea of collaborative filtering is endangered.

In his research, Cosley shows that users are influenced by the existing recommendation when rating an item (Cosley et al., 2003, p. 591). As the reader might have guessed, this does not mean we should give up the entire idea of collaborative filtering. First, Cosley’s results shows
that people are *influenced by*, not controlled by the existing rating. Second, his research was done on rating movies. As stated in the article, we cannot necessarily apply the results to other items, such as for example computer equipment. To sum up, when rating movies, Cosley shows that people are influenced by other users’ ratings. However, in order to determine whether these results apply for items outside the realm of movies, further research is needed. And then, of course, there are solutions. The first solution is a bit obvious: Hide the existing rating. If you cannot view other user’s opinions, they will not affect you. The second solution lies in the rating scale itself. This, however, is a subject which requires a chapter of its own.

In my research, I will focus on whether any of the cases offer the user the opportunity of to hide existing recommendations, and if I can find any other possible solutions to the problem of conformity.

### 2.4.5 Rating Scales

This section will present rating scales. I will start by presenting theory on scale granularity, then move a bit outside the world of computers and present theory on rating scales drawn from survey research. Finally, I will give a presentation on the usage of multiple rating scales.

When implementing a scale in a social navigation system, the *scale granularity* is essential. We can distinguish between four different solutions. The first solution is the *open ended* option. The user is asked to enter (as text) an example of an item (s)he likes, i.e. being asked to enter the name of one favourite artist. The second alternative is using a *Likert scale*. In this scenario, the user is asked to rate an item on a scale ranging from i.e. 1 to 5, where 1 could represent “strongly disagree”, 5 could represent “strongly agree” and the options in between various degrees of disagreement and agreement. A variation of the Likert scale is the use of a rating bar. This consists of a bar where each end is marked with i.e. “strongly disagree” and “strongly agree”. The user can then click anywhere on the bar in order to express an opinion. In this case, the user doesn’t have to rate an item on a fixed numeric scale (however, the system logic must translate the click into a numeric value). The third type of scales are *binary scales*, which, as the word implies, consists of an either/or choice. An example of this is to give an item the “thumbs up” or “thumbs down”. The last variation is *hybrid rating systems*, which combines two or more of the rating systems described above. An example of this is the “Britney Spears” button in this chapter’s initial quote. Here, the user is given the option of
either (1) rate the item on a scale (Likert scale) or (2) throw the item in the trash can (binary scale).

The challenge when it comes to rating scales and scale granularity is which one or ones to choose, which, as the reader might expect, is not simply picking the “best” one of the above. By looking at existing research, I have identified three different criteria which should be taken into consideration when creating a scale on which items are rated. First, it is important to design an easy and engaging process that keeps users from getting bored or frustrated when rating (Swearingen and Sinha, 2002). A combination of rating scales is a possible solution for this. A simple Likert scale can become repetitive and boring (Swearingen and Sinha, 2002). Second, Cosley et al. states that users prefer finer-grained rating scales (Cosley et al., 2003, p. 588), hence giving a small blow to the binary scale. The third criterion somewhat encompasses the two first, and states that the scale should allow the users to make meaningful distinctions (Cosley et al., 2003, p. 591). For example, rating a song on a scale from one to five should seem meaningful to the user, whereas rating a song either “good” or “bad” might for the user seem a too rough-grained choice.

The most commonly used rating scale in a social navigation context is the Likert scale. However, in the literature of social navigation, detailed studies of the scales themselves seem to be hard to come by. In order to build a theoretical framework in which Likert scales can be studied, I have chosen to look at literature on the use of rating scales in research surveys. Applying literature on rating scales in research surveys to social navigation rating systems should not constitute a problem, since the scales themselves (the Likert scale) are the same in both scenarios.

There are several ways in which a rating scale can be biased. In their article “Rating the Rating Scales”, Friendman and Amoo presents a synthesized overview of common pitfalls (Friendman and Amoo, 1999).

1. *Equal interval*. On a scale, category descriptions should have equal intervals. Take for instance the scale “terrible”, “horrible”, “awful”, “fair”, “slightly good”, “all right” and “reasonably good”. The perceived psychological difference between “awful” and “fair” would not be the same as between “fair” and “slightly good”. Further, different people perceive the same word or phrase differently. For instance, Friedman and
Amoo refers to a study done by Bradburn and Miles in which respondents were asked to define the words “very often”, “pretty often” and “not too often” with respect to days per month. The study concluded that there was a great deal of variability in how these phrases were interpreted.

2. *The number of points.* The number of points of which a scale should consist could be a bit tricky, but very important. The rating scale itself can affect the results of a survey (Amoo and Friedman, 2001). Cosley above states that a scale should enable the user to give a meaningful distinction. What is a meaningful distinction? First, the scale must be adapted to what is actually being rated. Second, the more scale points used the more reliable the scale, as long as the number of points is within what the subjects can handle. Friedman and Amoo writes, based on several experiments done by other researchers, that if one is interested in individual behaviour (which is probably the most relevant in social navigation systems), one should consider using a scale anywhere from 5 to 11 points.

3. *Assumption made by the question.* The question, and hence results, can be biased due to an assumption made by the question. For instance, asking a user “How good did you think this product was?” assumes the user actually thought it was good.

4. *Forced choice.* A forced choice rating scale, that is when there is no “I don’t know” option, can force users with no opinion on the subject to choose an option more or less randomly hence creating biased data. In surveys where there exists a chance of the participants not actually having an opinion or knowing something about a subject, they should be given the option of answering “I don’t know”, “no opinion” or the like. In the case of social navigation, rating an item is (as far as I’ve seen) optional. Hence, there is no forced choice and the problem solves itself. In any case, in a social navigation system, the user should not be forced to rate an item.

5. *Unbalanced rating scales.* Rating scales should be balanced, with an equal number of favourable and unfavourable choices. For example, the scale “excellent”, “very good”, “good”, “fair” and “poor” is not balanced, there are three favourable choices and only one unfavourable. Such a scale will easily result in too favourable results.

6. *Performance versus improvement scales.* When measuring subjects’ overall feeling towards an item, expectations questions provide more accurate results than a typical performance rating. An expectation question could for instance be “Overall, compared with what you expected, how would you rate…”. A performance question could be
“Overall, how would you rate…” In the case of social navigation systems, this could be very central, since rating the overall opinion of an item is very commonly used.

An important second aspect of rating scales concerns the number of different scales on which an item is rated. The most common choice of CF systems today is probably the use of one scale. For some items (i.e. songs) one scale might be satisfactory. But, say for example you’re rating a laptop, is rating from “good” to “bad” on a scale from 1 to 5 really a representative method of describing a computer? As Malaga points out, using multiple scales is a trade-off between number of categories and ease of use (Malaga, 2001, p. 413). Naturally, having multiple scales requires more of the user. On the other side, having just one scale might give an inaccurate description of an item. Unintentionally supporting Malaga, Swearingen and Sinha shows, through their research on the interaction between users and recommender systems, that users don’t mind giving a little more input in order to receive more accurate predictions (Swearingen and Sinha, 2002), thus supporting the case for more than one category. An additional solution found outside the rating scale itself, and mentioned in the theory section, is the use of written reviews.

My research will focus on the two following areas of the rating scales: First whether the scales fall into one of the six pitfalls described above. Second, I will look into the use, or lack of use, of multiple scales.

2.4.6 Confidence
If you receive a recommendation from a close friend, you would most likely put great trust in this recommendation. If you received the same recommendation from a stranger on the street, it is highly likely you’d put a lot less trust in the recommendation. This has to do with confidence. You are a lot more confident in your friend’s recommendation than you are in a stranger’s recommendation. If you put your trust in a CF system’s recommendation, you put your decision in the hands of a bunch of strangers. Harper, in his article on how information gains relevance, states that the context of the information (i.e. who wrote it? Is the web site reliable?) is of importance to the extent where using the Internet as an information source can be useless unless the person looking for information knows who is behind the source (Harper, 2003, p. 349-352).
The issue of confidence in CF environments varies depending on the item in question. Take for example a system that recommends songs and lets you play them. It doesn’t really matter much to you if the song recommended wasn’t that good. You can just skip to the next song. On the other hand, if you’re booking a hotel for your long-awaited bungalow vacation to Thailand based on an online recommender system, the issue of confidence becomes of great importance.

By looking at existing research, I have identified two somewhat similar solutions to the challenge of confidence. First, in a recommender system where confidence is an issue, why an item is recommended should be specified (Konstan and Riedl, 2003, p 72). In our hotel example, possible solutions could be to add pictures taken by the users and include the option of writing reviews. By doing this, confidence in the rating increases, since the user is specifically shown why this item is rated as it is.

Somewhat similar, Swearingen and Sinha has shown that transparency of system logic strongly affects the level of user trust. They state that [a] good CF algorithm that generates accurate recommendations is not enough to constitute a useful system from the users’ perspective. The system needs to convey to the user its inner logic and why a particular recommendation is suitable for them (Swearingen and Sinha, 2002). As a simple example of how inner logic can be displayed, take a system which shows to the users a simple average rating of an item. The number of other users who rated the item should be displayed since an item rated high by several users will most likely convey higher confidence as opposed to an item rated high by few other users.

Finally, it should be mentioned that reputation management systems present a bit of a special case when it comes to confidence, due to the fact that humans (the users) are rated rather than non-human items. Since I will not focus on this variation of CF systems in my research, the dilemmas will only be presented briefly. Among the major dilemmas I’ve found are (1) the ease of changing your identity if your rating turns out low, (2) fake transactions between friends in order to shill the rating system, (3) little thought-through methods of calculating overall reputation causing a false presentation of trustworthiness and (4) humans change over time, which means the reputation management system should be able to handle change in behaviour. For further reading, I suggest Zacharia’s, Moukas’ and Maes’ article on
collaborative reputation mechanisms, from which the examples above are extracted (Zacharia, Moukas and Maes, 1999).

Finally, it should be noted that confidence is an issue transcending the entire CF system. For instance, vulnerability to shilling attacks affects confidence. However, in order to categorize social navigation system rating system challenges, this section isolates the confidence issue to why items are recommended.

Based on the above presentation, my research will explore the issue of confidence. This will be done by looking at how well a CF system displays its inner logic and explains why an item is recommended.

2.4.7 Fading
If the Internet had been about 30 years older, and you entered a site for popular music today, the album on top of the “most popular” list might have been the West Side Story soundtrack from 1962 (the album that had the most weeks as number one on the Billboard 200 list). Naturally, you would be on the lookout for something more recent. This example illustrates why most CF environments requires a mechanism for fading of information.

The fading-of-information challenge is well known in the field of social navigation (Dieberger, 1997; Wexelblat and Maes, 1999, pp. 271; Dieberger et al., 2000, p. 44; Wexelblat, 2003, p. 227 and many more). The initial solution is simple: In most CF environments some sort of fading of ratings should be done, in order to fade out old or unwanted items.

I have had some trouble finding specific solutions to the fading problem in existing literature. The problem is recognized, but most literature seems to stop at the recognition and does not further elaborate possible solutions. However, there are a few well-known examples employed by many web sites. Amongst the possible solutions is the option of displaying items divided into different slices of time, such as displaying the most popular last month or all-time most popular. A similar solution is to sort items based both on age (how long since they were inserted into the CF system) and rating.
Reputation management systems were mentioned in the previous chapter as a special case when it comes to confidence. This is also the case in the issue of fading. Since humans can change, as opposed to say one specific movie, fading comes with extra complexities. Once a user’s profile has been established within the system, it becomes difficult to change one’s preferences (Burke, 2002, p. 338). As mentioned before, reputation management systems will not be a focus in this work, and I again refer the curious reader to Zacharia’s, Moukas’ and Maes’ article which discusses this issue (Zacharia, Moukas and Maes, 1999).

I have had trouble finding specific solutions to how the fading problem could be solved. A few are mentioned above, but these are mainly well-known solutions you can find on several existing web sites. In my research, I will therefore look at how the fading problem is solved (if solved at all), and be on the lookout for new solutions.

2.4.8 Privacy
Collaborative filtering systems gather data from the users of these systems, and hence, privacy becomes an issue. The challenge of privacy varies from system to system. In the case where a CF system gathers data about your actions within the system, i.e. your purchasing history, privacy can easily become a concern for the users. Your purchasing data can be tracked directly to you as a person, since you’re probably registered with at least a name and an address as a user of the system. On the other hand, CF systems where you anonymously vote on an item or in the case where you write a written review, privacy is less of an issue. In the first case, your vote cannot be traced back to you as a person. In the second scenario, it is you yourself who consciously made the review available to other users, and hence your privacy is the matter of your own choice.

Svensson et al. shows, in their study of an online system for food recipes, that as long as there is a perceived benefit, few of their test users minded being logged (Svensson et al., 2005). On the other hand, giving the user the option of adjusting their level of privacy within the CF system would give those users who are concerned about their privacy the possibility of not exposing themselves to a higher degree than wanted.

Due to the varying nature of CF systems, there are no clear-cut rules of how this challenge should be handled. Privacy is an important issue within social navigation, but in this work,
privacy will not be further elaborated. The reason for this is that I will be looking at how reliable social navigation rating techniques are as an aid in making your choices, not how reliable they are from a privacy point of view.

2.5 Challenges of User Reviews
Written reviews are prone to several of the same weaknesses and solutions as found when rating items on a scale. I have not found any examples of weaknesses in written reviews radically different from the weaknesses described in the preceding section. This section will give a description of shilling attacks, barrier-to-entry and confidence in the light of written reviews.

As with rating systems which use scales, written reviews are open to shilling attacks, both push and nuke attacks. As with scales, items with few existing reviews are more exposed than items with several reviews. Some of the same solutions to avoid shilling attacks on rating scales can be employed in the case of written reviews. Technical barriers to avoid bot attacks and massive postings of reviews from a single source can rather easily be put in place. A second possible technique, reflecting the trusted-source rating on scales, is the option of using expert reviews for new items. Algorithmic detection of written review shilling attacks is possible but more complicated than on scales, since computers have a hard time understanding the actual content of a piece of text.

The equivalent of the barrier-to-entry challenge for written reviews would be the simple lack of reviews. Considering that writing a review is more time and resource consuming for the user, this challenge could be more severe for written reviews than ratings on a scale. One possible technique which could be employed in order to alleviate this challenge would be to reward the user who writes a review, in the same way as described with ratings on a scale.

The last challenge is confidence. As reviewing an item requires considerably more knowledge from the user who writes the review, as opposed to simply rating one’s opinion on a scale, confidence becomes a more important issue. Who is behind this review? Does the person have the necessary knowledge to write a trustworthy review? One possible solution is found in the distinction between expert reviews and other reviews described in the theory section. Reviews from known experts or trusted individuals could be marked as such, in order to build trust in
the review. This does not however solve any problems for the remaining reviews. Related to
the challenge of confidence are senseless reviews, that is, reviews which do not actually say
anything useful for other users about an item. For instance, “this movie was horrible” does not
have any value to the reader, because it does not say why the movie was horrible. Similar to
transparency of logic and displaying why something is recommended in ratings on scales, in
order to gain confidence, the claims given in user written reviews should contain an
explanation in order to avoid more or less senseless reviews.

User reviews are exposed to several of the same challenges as ratings on a scale. Due to this
similarity, the focus of my research will be on confidence, looking closer at how useful
written reviews are to the users.
3. Method

There are three kinds of lies: lies, damned lies and statistics.

- B. Disraeli

This chapter describes the methods used in my research. I will first start by giving a theoretical introduction to the research methods applied in my work, then move on and describe how the cases were selected and finally present the specific methods used in the studying of the cases.

3.1 Method Theory

3.1.1 Collecting Data

There exist several different approaches when collecting data. The selection of method for data collecting depends on which of the methods we think would be the most valuable in answering the question at hand. The following sections will present the theory behind the data collecting methods used in this thesis.

Document Analysis

Document analysis is traditionally used about printed text, but due to technical advances, this sort of information is today found in various types of media. Documents can include statistics, archives, public documents (i.e. journals and protocols), private documents (i.e. diaries), literature (i.e. fiction and non-fiction), newspapers, video and audio (Patel and Davidson, 1999, p. 48-49). Document analysis can help answer questions about factual occurrences and events, such as for instance how a newspaper article can describe a specific situation. It is however important to remain critical of the documents analyzed. Who is behind this document and what are their intentions? Does this document represent an accurate description of reality, or has time and/or external influence reduced the document’s validity? Is this document a fake? One can distinguish between primary and secondary sources (Patel and Davidson, 1999, p. 50). A primary source is the object itself, i.e. a historical event. A secondary source is a source one “step” away from the original, i.e. a newspaper article about
the historical event. Finally, it is important to not only select documents which support our own ideas. By doing so, the picture will not be complete, or in the worst case, wrong. It is important to look at a case from several angles, and try to seek out information which possibly counters our results.

**Questionnaires and Rating Scales**
Rating scales are frequently used in questionnaires. Rating scales can help reveal an individual’s attitude towards a specific item (Patel and Davidson, 1999, p. 62). One of the most commonly used rating scales is the Likert scale, in which an individual is asked to rate an item on a scale from (for instance) 1 to 5, where 1 could mean “strongly agree” and 5 could mean “strongly disagree”. How rating scales are formed can have a great effect on the results, and if done incorrectly, could bias the results (Friedman and Amoo, 1999). Somewhat unusually, since rating scales are a property of recommendation systems, the subject of how rating scales should be created has already been discussed. Since I see no reason to repeat what I’ve already written, this subject will not be further elaborated. The forgetful reader may take a look at the preceding chapter on rating scales in the theory section.

**3.1.2 Analyzing the Data**
When data has been collected, it must be analyzed and interpreted in order to be able to give an answer to the questions asked. This section will introduce two general methods of analyzing data, namely *quantitative* and *qualitative* methods.

**Quantitative Methods**
Qualitative data is analyzed using statistics. Statistics is a science on its own, and this section will only give a brief description of some of the basic terms and methods. Our first concern is that we would like the statistical results to apply to an entire population of items. However, gathering data from an entire population is usually not possible. For instance, if we would like to find out what percent of a population would vote for a specific party, asking every single inhabitant of a country would be a rather tedious task. So, what we do is make a representative selection of the population, or in other words, we have to select a group of people who reflect the entire population in order to be able to generalize our results so we can safely claim they apply to the entire group we are studying. Representative selection is absolutely crucial for the result. For instance, if you were to select people for an inquiry into
political sympathies by asking random individuals exiting a very exclusive restaurant, the results would be biased, since this group of people is not representative for the entire population (the results would probably lean too much to the right). Even if you kept on asking hundreds of thousands of people, your results would still not be correct. There are two different ways of making a representative selection (Patel and Davidson, 1999, pp. 79-81). The first is simply to pick a random selection of individuals from the population. Say you have 10,000 individuals for which you want your results to apply, you could simply pick say 100 of these at random. This method has some limitations, however which can be helped using stratification. Say your population is 30% men and 70% women, and we want our results to reflect this distribution. What we do is to divide the 10,000 individuals into two strata, men and women, then randomly select individuals from these two groups proportional to the size of the strata. Thus, our selection accurately reflects the gender distribution of the population.

In a quantitative study, the property studied is called a variable. Each variable is measured on a scale. We can distinguish between four kinds of scales (Patel and Davidson, 1999, pp. 80-81). A nominal scale measures in categories, i.e. gender. An ordinal scale measures in range, but without specifying the distance between the points measured, i.e. first through third place in a sports competition. An interval scale measures within a range with a known distance between the points, i.e. measuring temperatures. The last scale is called the quota scale, which measures with a known distance between the points and with an absolute ending point, i.e. number of children in a family. A second property of variables is whether they are discrete or continuous (Patel and Davidson, 1999, p. 81). A discrete variable can only contain a certain set of values (usually whole numbers), i.e. the number of children in a family. A continuous variable can have all values within an interval, and it is the granularity of the scale which decides how accurate the distinctions are. For instance, temperature could me measured in whole degrees, tenth of a degree, hundredths of a degree and so on.

The next step is to present the data in a meaningful way. This could range from simple calculations of average to complex statistical analysis. The method(s) chosen depends on which questions one seeks to answer. The results are most commonly presented as either a table of numeric values or a chart, such as bar chart or pie chart. This work relies on simple statistical methods which require no further explanation, such as the calculation of averages and distribution of measured values.
Qualitative Methods
The purpose of qualitative methods is usually to gain a deeper understanding of a phenomenon which cannot easily be revealed using quantitative methods (Patel and Davidson, 1999, p. 88). For instance, statistics can explain how many people are likely to vote for a party, but cannot readily explain why people chose to vote for a specific party. Qualitative analysis can be performed in a myriad ways, and is not as easily definable as quantitative methods. One of these methods, used in this thesis, is heuristic evaluation. This method uses a set of guidelines or general principles, and based on these evaluates an object (Dix et al., 2004, pp. 324-326). Let’s illustrate with a simple example. Say you have figured out that using the standard blue underlined text in links on a web site causes increased usability, you define the following principle: “Web sites should use the commonly known blue underlined text in links”. This is your general rule. By looking at how an existing web site colors their linked text, you do a heuristic evaluation based on a general rule.

Finally, the author would like to note that heuristic evaluation is not necessarily qualitative, however it is mostly used qualitatively in this work. Simply put, there is no rule stopping us from creating a heuristic evaluation principle which requires a statistical analysis.

3.2 Selecting the Cases
All my cases are existing web sites. When selecting the cases, I used the below criteria:

1. All the web sites must contain a product recommendation system. Recommendation must be done (1) through some sort of collaborative filtering system and (2) with the ability of users to write reviews.
2. The product type recommended must be computer equipment. The reason for this is twofold. First, most research on collaborative filtering systems seems to focus on items with a high degree of subjective properties. For example, when an individual rates a movie or a song, there is a higher degree of subjectivity involved as compared to computer equipment. In choosing computer equipment, my research becomes more unique, since it focuses on items with a higher degree of objective properties. The second reason, which explains why I didn’t choose a different set of technical items (i.e garden tools), is simply that since this is a computer science master thesis, the
author is somewhat convinced that he has some knowledge in the field of computers, and therefore is able to make clearer judgements on the data collected if it should be required.

3. The web sites selected should be Norwegian. There are two reasons for this. The first reason is simply that there are so many product recommendation sites out there, that I’ll have to narrow it down one way or the other. Second, due to the fact that products vary across countries (i.e. a product with the same model number can vary in technical specifications from country to country), comparison between specific products become problematic. However, since my focus is on the collaborative filtering mechanisms implemented on the selected web pages, and the fact that I’m looking at computer equipment, should make my results comparable to other, similar web sites in other countries.

3.3 Studying the Cases
In the theoretical section of this thesis, several weaknesses of social navigation rating techniques were pointed out. Due to the number of challenges and their varying nature, I have chosen an approach in which several small studies have been done, both quantitative and qualitative. In this section, a description of how the research was done is presented. In short, two specific methods have been used: Heuristic evaluation of the web pages themselves and statistical analysis of data gathered from the web sites. If the web site itself could not answer the question posed by the heuristic evaluation, e-mails were sent to the companies behind the sites.

The cases selected were studied in relation to the weaknesses of collaborative filtering systems presented in the theory section of this thesis. My primary goal is to give an answer to the following three questions (as stated in the introduction):

1. Through the study of social navigation theory and existing research done on social navigation systems, a set of known weaknesses can be compiled. Through an analysis of existing web-based recommendation systems, how exposed are these to the known weaknesses?
2. By analyzing web-based recommendation systems, are there any weaknesses not identified through question 1?
3. By analyzing web-based recommendation systems, are there any strengths not identified through question 1?

Question two and three are pretty straightforward, but question one requires a more detailed description. I have devised a method for each of the known weaknesses described in the theoretical section, with the purpose of giving a degree of measure of exposure to the identified weakness. A separate investigation was done for each method in each case. The methods used are described in the following sections. As previously noted, the challenges of collaborative filtering algorithms and privacy will not be subjects of my research, and are hence not further elaborated in the following sections.

3.3.1 Shilling attacks
In my research on shilling attacks, I will focus on the degree to which it is possible to expose a CF system to false ratings. Based on the theory section, the following aspects of my cases will be studied:

- Missing ratings. As described in the theoretical section of this thesis, CF systems with few or missing ratings are a lot more vulnerable to shilling attacks. Based on this, I have chosen the following approach:

1. I will focus on four important product categories: Laptops, desktops, inkjet printers and LCD monitors. I have chosen these categories since they are what a common home or office computer system would consist of.
2. All ratings for products in these categories were recorded. This was done within as short a time period as possible (four days), in order to avoid any differences between cases in the data recorded since the web sites are open to new ratings as data is recorded.
3. The following variables were calculated: The number of rated and unrated items, the average number of ratings on rated items and the distribution of ratings. The lower the average numbers of ratings and the lower the average number of ratings per rated item, the more susceptible the items are to a shilling attack.
This method has some limitations. Ideally, the basis for my calculations should be the entire databases of my cases. Since this was not an option – it would require four competing companies to give the author access to their databases – I narrowed my selection of products down to what a common home or office computer system would consist of. However, in order to get the complete picture, the entire database should be used.

- **Counter-measures.** Look into which counter measures are employed to avoid shilling attacks, and identify whether the following three counter-measures identified in the theory section are used:

  1. Technical limitations, such as disabling massive voting by clicking like crazy.
  2. Special protection of new items.
  3. Built-in functions in rating-algorithms to track down false rating patterns.

### 3.3.2 Barrier to Entry

In my study of the challenge of how to enable new and unrated items to receive ratings, I will have a two-tier approach:

1. Based on the rating data gathered in the study of shilling attacks, analyze how extensive the barrier-to-entry problem is.
2. See if any of the following mechanisms for alleviating the barrier-to-entry challenge are employed:
   a. Offer users a reward for rating.
   b. Passive data collecting.
   c. Default rating for new items.
   d. Hiding the rating of newly added items.
   e. Increased exposure of new items.

### 3.3.3 Conformity

On the issue of conformity, I will not look into how susceptible the users of my cases are to ratings done by other users when they themselves rate an item. This is simply because this
study would require a thesis in itself. When it comes to conformity, I will simply look at whether my cases have an option of hiding the existing recommendations when rating, or any similar solution. (Note the difference between the option of hiding ratings for *new* items and the option of hiding ratings for *all* items).

### 3.3.4 Rating Scales

In the case of scale granularity I will have a threefold approach. First, I will evaluate the use of rating scales according to the six criteria identified in the theory section. These are:

1. *Equal interval.* I will look into whether the scales used have equal intervals.
2. *The number of points.* I will look into whether the cases use a scale ranging from 5 to 11 intervals.
3. *Assumption made by the question.* I will look into how the rating question is asked.
4. *Forced choice.* I will look into whether any of the cases forces the user to give a rating.
5. *Unbalanced rating scales.* I will look into whether the scales employed are balanced.
6. *Performance versus improvement scales.* I will look into which (if any) of these scales are used.

Second, I will look into the usage of multiple scales. According to theory, the usage of multiple rating scales would be useful, since computer products are complex and not easily rated on a single scale. If several scales are found to be used, they will all be evaluated according to the criteria above.

Third, I am going to perform two sets of statistical analyses. First, I am going to compare the rating of the same item across the four cases using the data collected on the statistical analysis of shilling attacks. If ratings work according to the social navigation ideal, these ratings should be equal (or almost equal) across the cases. Second, I am going to calculate the distribution of the ratings. Ideally, ratings should be equally spread out. If the majority of ratings given are centred on a specific area of the scale (i.e. a majority of the ratings given are very high), one could question both the scale itself and the recommendation. If the ratings are centred on an area of the scale, maybe the scale itself is not suitable for rating the items. Secondly, the case’s role as a recommendation system is less trustworthy since the average
rating does not in practice utilize the entire scale and hence does not accurately reflect the item’s rating relative to the scale and/or other items.

### 3.3.5 Confidence
The cases I am looking at are product recommendation systems. The information displayed on these web sites is meant as an aid for potential computer software and hardware customers. Hence, confidence is of great importance. On the issue of confidence, I will analyze and categorize which measures are employed, other than the ratings themselves, by the cases in order to build confidence in the product recommendations, and whether the systems displays the inner logic of the CF system used to rate the items.

### 3.3.6 Fading
Fading, in the world of recommendation systems for computer software and hardware, as I see it, is the CF systems’ ability to fade away old products. For instance, if you’re in an alphabetically sorted list of operating systems today (2007) and find Windows Me before Windows XP, fading has failed. In my research, I will look at how the recommender systems fade away products which are getting old and less interesting.

### 3.3.7 Written Reviews
As already noted, written reviews are prone to several of the same weaknesses as ratings on a scale. However, weakness to shilling attacks and barrier-to-entry is largely covered by my research on the rating scales, and I hence chose a different approach for the written reviews, namely evaluating how useful a group of people found the written reviews to be. Since evaluating written reviews is a very difficult task, my approach requires some elaboration.

One approach in rating written reviews is to assess the contents on a positive to negative scale and then analyze these results in much the same way as I did with the ratings on a scale. This was attempted by Godes and Mayzlin in a paper analyzing Internet newsgroups’ text (Godes and Mayzlin, 2004, p. 555-556). Their approach was to ask two independent readers to evaluate written reviews of TV shows on a scale ranging from “Positive” to “Irrelevant”. However, only 57% of the texts received identical evaluations from both reviewers. They
conclude that accurate content analysis is extremely difficult due to its subjective nature, and hence this approach will not be used in this work.

A second approach is to evaluate the context of the review according to a set of absolute criteria, such as counting the number of substantiated statements about the item (i.e. “I did not like this computer, because there was a driver problem with the wireless adapter”) and the number of unsubstantiated statements (i.e. “This computer comes with Windows Vista, and since I don’t like Bill Gates, I hate the computer.”). By doing this, one could get an impression of the credibility of the reviews. This approach was briefly attempted but quickly dropped. It turned out to be a quagmire of textual analysis mainly because the text itself is too complex to be analyzed even by absolute criteria like the ones above.

The following approach was finally chosen. Three groups of 12 product reviews were randomly selected from the cases. The products selected were all within the categories used when studying shilling attacks in order to focus the study on the same group of products. Three persons were then asked to evaluate each group, totalling nine participants. Through this combination of individuals and reviews I get (1) several individuals’ response on the same reviews and (2) a certain depth in reviews since a total of 36 different reviews were evaluated. The survey was done through a web based form, on which links to the reviews themselves were placed and the reviews were read directly from the web pages on which they were found. The participants were asked to put themselves in a context in which they were considering buying the products in question, and then evaluate the written review on a scale from 1 to 5 specifying how useful the review was to them in this context. Additional questions like age, gender, experience with computers and experience with recommendation sites were also recorded. Only participants with a certain level of experience with computers were selected. The reason for this is that some of the reviews can be a bit technical. If the participant does not understand the review, the participant does not know what s(he) is rating, and hence the result is biased. The first group of three participants were in addition used as a preliminary survey. In this approach, I give in to the subjective nature of rating a review. The participants were specifically asked to rate degree of usefulness/lack of usefulness for themselves. The reason for this is twofold. First, what is useful to me may not be useful to you and vice versa. Hence, objective usefulness in this context is not possible. Second, my goal with this work is to assess the reliability and usefulness of recommender systems. Evaluating individuals’ subjective opinion about a review can thus give an indication of how useful these
written reviews actually are and hence be used as a base for evaluating the implementation of these social navigation techniques in the four cases.
4. Introducing the Cases
This chapter gives a short introduction to the cases studied in this work. Each introduction starts with a short description of the web site, then moves on to give an example of how one product is presented on the site. Many of the sites offer services outside product recommendations and user ratings/reviews, but since this thesis focuses on social navigation rating techniques, extra services will only be mentioned very briefly.

4.1 Akam
Akam (produkt.akam.no) is the result of a merger between Digitalkamera.no and Videokamera.no, and is associated with the web site Hardware.no. They offer a wide range of services, from expert recommendations and reviews, second-hand sales, forums and product tests, price comparison from different web shops, user ratings and reviews of the web shops themselves and users’ product ratings and reviews. The site’s main focus is on photo and video, but in their product guide section on which I focus, they offer product information on computer hardware, software, mobile phones, hi-fi and TVs.

Within their computer hardware and software product guide, they offer a very wide range of product descriptions, ranging from desktop computers to single cables. Below is a screenshot of how a single product is presented on the screen.
You can browse between six categories of product information. The first category (displayed in the screenshot above) labelled “Forside” (“front page”) contains basic technical information about the product and an image. The second category, “Erfaringer” (“experiences”), contains an overview of other users’ ratings and reviews of the product, and the option of creating your own rating/review. The third category, “Detaljer” (“details”), offers a detailed technical specification of the product. The forth category, “Bilder” (“pictures”), shows various pictures of the products. The fifth category, “Priser” (“prices”), shows a list of web shops and the prices of this product at the specific web shop, sorted by cheapest first. The last category, “Lenker” (“links”), displays a list of various links related to the product, such as manufacturer’s website and articles about this product. A trimmed version of links and experiences, and the entire list of prices are shown on the front page, below what is seen in the screenshot above.
4.2 DinPris

DinPris (www.dinpris.no) is, according to the “About Us” section on their site, an *objective and independent price comparison service making it simpler for the users to shop online.* The site focuses mainly on user created product and services information and price comparison.

The site offers the following services: Comparison between two or more products in the same category (i.e. two mobile phones), recommending products to your friends, price comparison from different web shops, user ratings and reviews of the web shops themselves and users’ product ratings and reviews. DinPris offers information on both products and services, ranging from rental cars to furniture. Within their computer hardware and software section, the site offers a wide range of product descriptions, ranging from desktop computers to single cables. Below is a screenshot of how a single product is presented.

You can browse between four categories of product information (lower bar in picture above). The screenshot shown above will always remain as it is, category information is shown in addition. As seen on the picture, an average rating, a short product description and cheapest/most expensive price is shown. The category “Sammenlign” (“compare”) shows a list of web shops and the prices of this product at the specific web shop, sorted by cheapest first. When clicking on “Teknisk spesifikasjon” (“technical specification”), a short overview of technical product information is displayed. The third category, “Innlegg” (“reviews”), shows reviews and ratings given by other users. The last category, “Artikler & Nettsider” (“articles & web pages”), shows a list of links about or related to the product and/or the product category (i.e. guide to setting up a wireless network).
4.3 Kelkoo
Kelkoo (www.kelkoo.no) has a twofold purpose. First, it aims at helping other users shopping online and finding the products they’re looking for. Second it has a commercial aspect in that it helps shops and brands in increasing their sales through increased product exposure. The site was established in 1999, and is available in ten countries (in my study, I only used the Norwegian site). Kelkoo became a part of Yahoo! in 2004. The site is clearly more commercial compared to the other cases, mainly through the use of ads for products and web shops.

The site offers the following services: purchasing guides, price comparison from different web shops, user ratings and reviews of the web shops themselves and users’ product ratings and reviews. Kelkoo offers information on a wide range of products and services, ranging from plane tickets to coffee machines. Within the computer hardware and software section, as with the two above web sites, the product spectrum ranges from desktop computers to single cables. Below is a screenshot of how a single product is presented.

![Picture 7: Presentation of a product on the Kelkoo web site.](image)

Kelkoo offers very limited technical product information. The above is all there is. By clicking on the linked product name, you are taken to the web shop with the lowest price for this product. The link “Vær den første til å skrive et norsk innlegg” (“be the first to write a review in Norwegian”) enables you to write a product review. The link “Vurder dette produktet” (“rate this product”) enables you to rate the product on a scale from 1 to 5. The average rating is shown as a row of stars above. In the upper right corner, you can see the rating of the web shop which has the lowest price on this product.
4.4 Komplett
Komplett (www.komplett.no) differs from the other cases in that it is a web shop. Komplett.no is a part of Komplett ASA which has web shops in 10 different countries. In this study, only the Norwegian site was used. The site offers its registered users the option of rating and reviewing products within their product spectrum, thus enabling potential buyers of a product to see other buyers’ experiences.

The sites’ main focus is naturally selling products, but in addition to this offers user-created ratings and reviews and a “people who bought this item also bought” feature. The site focuses on consumer electronics, ranging from kitchen equipment to computer equipment. Below is an example of how a single product is presented.
At first, some very basic product information and the product’s average rating is displayed. Below this again you find extended product information. Further down (not shown in the picture) you find the users’ ratings and reviews and finally other products associated with this product (such as for instance wireless adapters for the computer). By clicking on “Utvidet info” (“extended information”) a page with extensive product information is displayed. By clicking

---

4 One could question whether this feature works well on a web shop which sells electronic equipment. As seen in the picture above, it seems like people who bought this wireless router also bought cooling paste. This recommendation does not have much value.
on “Tester” (“tests”, faded in the picture above) one or several tests of the product are displayed (not all products, as in this case, have a test associated with it). The tab “Produsentlinker” (“manufacturer’s links”) gives you a list of external links to the manufacturer’s website. Finally, by clicking “Tips en venn” (“tell a friend”) you’re presented with a form enabling you to tip a friend about this product.
5. Research Findings
This chapter presents the findings of my research, and some brief comments on the results. A general discussion is found in the next chapter.

5.1 Shilling Attacks
My first approach was to do a statistical analysis of item ratings. A total of 3511 items were surveyed.

Rated versus not rated items
The table below shows how many items had received one or more rating(s) and which items had received no ratings.

<table>
<thead>
<tr>
<th>Case</th>
<th>Total number of items</th>
<th>Items rated</th>
<th>Items not rated</th>
<th>Items rated, in percent</th>
<th>Items not rated, in percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akam</td>
<td>1755</td>
<td>237</td>
<td>1518</td>
<td>13.50 %</td>
<td>86.50 %</td>
</tr>
<tr>
<td>DinPris</td>
<td>766</td>
<td>4</td>
<td>762</td>
<td>.52 %</td>
<td>99.48 %</td>
</tr>
<tr>
<td>Kelkoo</td>
<td>702</td>
<td>355</td>
<td>347</td>
<td>50.57 %</td>
<td>49.43 %</td>
</tr>
<tr>
<td>Komplett</td>
<td>288</td>
<td>42</td>
<td>246</td>
<td>14.58 %</td>
<td>85.42 %</td>
</tr>
<tr>
<td>Total:</td>
<td>3511</td>
<td>638</td>
<td>2873</td>
<td>19.79 %</td>
<td>80.21 %</td>
</tr>
</tbody>
</table>

Table 1: Rated versus not rated items.

One can divide these results into three categories. In the case of DinPris, only .52 percent of the items recorded were rated, actually making the author responsible for 25 percent of all the ratings (!). The second category is Akam and Komplett, which has an average of 13.5 percent and 14.58 percent rated items. Even though, compared to DinPris, these sites have 2 596 percent and 2 803 percent higher number of rated items respectively, the numbers are still very low. Our last site, Kelkoo, has a stunning 9 725 percent higher number of rated items compared to DinPris, rendering about half the products rated by one or more user(s). Though the number of rated items is far higher relative to the other results, the numbers are still quite low.

Average number of ratings
The table below shows the number of rated items and the average number of ratings per rated item.
<table>
<thead>
<tr>
<th>Case</th>
<th>Total number of ratings</th>
<th>Total number of rated items</th>
<th>Ratings per item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akam</td>
<td>565</td>
<td>237</td>
<td>2.38</td>
</tr>
<tr>
<td>DinPris</td>
<td>4</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>Kelkoo</td>
<td>2612</td>
<td>355</td>
<td>7.36</td>
</tr>
<tr>
<td>Komplett</td>
<td>132</td>
<td>42</td>
<td>3.14</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>3313</strong></td>
<td><strong>638</strong></td>
<td><strong>5.19</strong></td>
</tr>
</tbody>
</table>

Table 2: Average number of ratings per rated item.

This table gives an indication of how easy it would be to perform a shilling attack on an already rated item. The higher the average number of ratings per item, the more difficult it is to perform a shilling attack. 1 is the lowest possible value. On average, if DinPris was exposed to a shilling attack, one would require *one* new fake rating in order to counter 100 percent of the existing ratings. For Akam and Komplett, one would require 2.38 and 3.14 ratings respectively, and 7.36 ratings for Kelkoo.

**Distribution of ratings**

Averages can hide differences. The graphs below show how ratings are distributed. The X-axis shows the number of ratings and the Y-axis the number of items with this number of ratings. In short, the higher the number of ratings and the higher the number of items with high ratings, the harder it is to perform a shilling attack. In the graph, this would mean many dots towards the upper right corner. On the opposite, the fewer items with low numbers of ratings, the easier it is to perform a shilling attack. In the graph, this would mean many dots in the lower left corner. The fourth graph shows the preceding three graphs on the same scale, in order to gain a valid comparison. (Due to insufficient data, DinPris is not included in these statistics).
Graph 1: Rating distribution at Akam.

Graph 2: Rating distribution at Kelkoo.
Graph 3: Rating distribution at Komplett

Graph 4: Combined rating distribution.

Both Akam and Kelkoo exhibit the same clear pattern, though Kelkoo is a bit better off than Akam. Most of the rated items are rated by very few. In the case of Akam, 49 percent of rated
items have only one rating. The item receiving the most ratings was rated by 26 users. In the case of Kelkoo, 24 percent of rated items have only one rating, and the item receiving the most ratings was rated by 79 users.

At first glance, Komplett looks a bit brighter. On the first graph, the number of ratings is rather high. However, and as the second graphs shows, there are in fact rather few items being rated by rather few people. No item is rated by more than five users, and only 16 products have this many ratings. As seen in the last graph, both Akam and Kelkoo have a better rating distribution than Komplett. In other words, Komplett exhibits a higher degree of vulnerability compared to Akam and Kelkoo when it comes to the probability of finding an item with a low number of ratings.

**Technical limitations**

Technical limitations can prevent against shilling attacks. The below table shows which technical protective measures were found in the four cases.

<table>
<thead>
<tr>
<th>Case</th>
<th>Measure</th>
</tr>
</thead>
</table>
| Akam    | • Distorted text (“messy” image with letters which bots have a hard time reading, but can be read by humans).  
           | • Manual review of rating.                                               |
| DinPris | • Manual reviewing of rating.                                           |
| Kelkoo  | • Must register an account before you can write reviews (not ratings on a scale). |
| Komplett| • Must be a registered customer.                                         |

Table 3: Technical prevention of shilling attacks.

**Protection of new items**

I have found no measures in any of the cases to specially protect new items.

**Seeding of new items**

None of my cases seed new items with a rating, neither an average rating nor expert rating. However, both Akam, DinPris and Komplett offer links to external reviews and tests of the items.

**5.2 Barrier to Entry**

The barrier-to-entry challenge is closely related to the number of items rates, and as seen in the previous chapter, the cases suffer rather heavily from the barrier-to-entry problem.
Kelkoo, with the highest number of rated items, had 50 percent of the items studied rated, whereas DinSide only had a mere .52 percent of the studied items rated.

But, which measures are taken in order to alleviate this problem? The table below shows the list of identified possible solutions for the barrier-to-entry problem and whether they were found in my cases or not.

<table>
<thead>
<tr>
<th>Case</th>
<th>Offer users a reward for rating</th>
<th>Passive data collecting</th>
<th>Default rating for new items</th>
<th>Hiding the rating of newly added items</th>
<th>Increased exposure of new items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akam</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DinPris</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Kelkoo</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Komplett</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4: Employment of barrier-to-entry countermeasures.

One of the cases, Akam, has *none* of the known alleviating techniques implemented. DinPris and Kelkoo are using passive data collecting, but in a rather limited manner. The number of users viewing an item is recorded, but they are not used as a basis for rating, but only for creating a “most popular” option of sorting items within each category. DinPris and Komplett implements increased exposure of new items. However, exposure is limited. In the case of DinPris, it is used as a sorting option within each category. In the case of Komplett, it is implemented as a textual list of items found at the bottom of the front page (nearly overlooked by the author).

In total, with five different techniques for alleviating the barrier-to-entry problem and four cases, only 4 of 20 possible implementations were found.

5.3 Conformity
This is probably the shortest chapter in this thesis. None of the cases offers the opportunity of hiding all existing ratings.

5.4 Rating Scales
Three of the cases use a single rating scale, whereas Akam uses four. All four scales in the case of Akam use the same scale interval and naming, and can hence be treated as one in most
of the examinations below. Kelkoo represents a bit of a special case, since it offers two different ways of rating an item. In addition to rating an item on a single scale, you can rate the item on several scales adapted to each product category. This is only possible in combination with writing a review, and requires the user to register a Yahoo! account. The ratings on Kelkoo will hereafter be referred to as Kelkoo 1 and Kelkoo 2, where Kelkoo 2 is the scale which is used in combination with the written review.

The table below shows the rating scales used. The numeric row to the left shows how many rating options each case has. Below is a picture of how the rating scales actually look.

<table>
<thead>
<tr>
<th>Case</th>
<th>Granularity</th>
<th>Akam</th>
<th>DinPris</th>
<th>Kelkoo 1</th>
<th>Kelkoo 2</th>
<th>Komplett</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Katastrofalt (Catastrophic)</td>
<td>No written scale. Rating is done by choosing between five rows of stars, starting from 1 ending at 5.</td>
<td>No written scale. Rating is done by moving the mouse over a row of stars, choosing between 1 and up to 5 stars.</td>
<td>Utmerket (Excellent)</td>
<td>Dårligst (Worst)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Svært dårlig (Very bad)</td>
<td></td>
<td></td>
<td>Veldig bra (Very good)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Dårlig (Bad)</td>
<td></td>
<td></td>
<td>Bra (Good)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Nokså dårlig (Pretty bad)</td>
<td></td>
<td></td>
<td>Middels (Medium)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Middels (Medium)</td>
<td></td>
<td></td>
<td>Dårlig (Bad)</td>
<td>Best (Best)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Nokså godt (Pretty good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Godt (Good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Svært godt (Very good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Strålende (Brilliant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Perfekt (Perfect)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Overview of rating scales.
Equal Interval
The scale used by Akam does not have equal intervals, though close. When judging whether a textual scale has equal intervals, it is hard to avoid a degree of subjectivity. But, in the author’s eyes, the distance between “catastrophic” and “very bad” is not equivalent to the distance between “brilliant” and “perfect”. On the other hand, Akam combines both a textual and numeric presentation of the scale (each textual label is numbered), making intervals more clear.

The case of DinPris and Kelkoo 1 poses a problem. They do not have a textual scale at all. The scale intervals are equal, since the interval between the choice of number of stars in this case is necessarily equal. On the other hand, we have no information of what the scales actually move from and to. I would assume most people think of one star as worse than five, but how bad is one star? How good is four stars? The Kelkoo 2 scale does not have equal
intervals. For instance, the distance between “good” and “very good” is shorter than the distance between “medium” and “bad”. The last case, Komplett, is not that far from DinPris and Kelkoo. The interval is again necessarily equal, but we are informed of what the scales are from and to.

**Number of Points**

Akam has scales ranging from 1 to 10. All the other cases have scales ranging from 1 to 5. It is worth noting that Akam displays the average rating of an item by using a 1 to 5 scale of stars, hence translating from four 1 to 10 point scales to one 1 to 5 point scale. To quickly conclude, all the cases use scales within the advised 1 to 11 range.

**Assumption made by Question**

The below table shows how the rating questions is asked including any instructions given to the user on how rating should be done.

<table>
<thead>
<tr>
<th>Case</th>
<th>Question(s)</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akam</td>
<td>[instructions] Karakterskalaen vår har ti trinn. Tenk deg godt om før du avgjør hvor du vil plassere karakteren. 1 og 10 skal representere ekstremitetene – tilnærmet komplett håpløshet, og tilnærmet perfeksjon. [questions] Til slutt noen spesifikke karakterer om hvor bra produktet er: Brukervennlighet Pris/ytelse Kvalitet Total</td>
<td>[instructions] Our rating scale has ten levels. Think thoroughly before making your choice. 1 and 10 represents the extremes – close to complete uselessness and almost perfect. [questions] Finally, some specific ratings on how good the product is: User friendliness Price/gain Quality Total</td>
</tr>
<tr>
<td>DinPris</td>
<td>[nothing]</td>
<td>[nothing]</td>
</tr>
<tr>
<td>Kelkoo 1</td>
<td>Vurder dette produktet Bruk musen til å peke på stjernene for å vurdere dette produktet Klik på stjernen for å lagre</td>
<td>Rate this product Use the mouse and point at the starts to rate this product. Click the star to store</td>
</tr>
<tr>
<td>Kelkoo 2</td>
<td>Vurder dette produktet</td>
<td>Rate this product</td>
</tr>
<tr>
<td>Komplett</td>
<td>Hvordan vurderer du dette produktet?</td>
<td>How would you rate this product?</td>
</tr>
</tbody>
</table>

Table 6: Rating questions.
Akam, though having the most extensive rating instructions, falls into the trap of asking “how good” the user thinks the product is. DinPris does a minimalist twist and asks nothing at all. Finally, Kelkoo 1, 2 and Komplett has a neutral formulation of their question by asking the user simply “to rate” the product.

**Forced Choice**

None of the cases forces the user to rate anything. It is worth noting however that Akam and Komplett forces the user to include a written review of the product in combination with the rating. DinPris and Kelkoo offers the option of doing either or both.

**Unbalanced Rating Scales**

The scale used by Akam is slightly unbalanced in a positive direction, in that five of the choices are positive, one is neutral and four are negative. The scale used by Kelkoo 2 is more unbalanced, three of the choices being positive, one neutral and one negative. The three remaining scales cannot be judged as being balanced or unbalanced, due to the lack of textual labelling.

**Performance versus Improvement Scales**

Except for DinPris, which asks no questions in combination with rating, *all* the cases use performance ratings. As described in the theory section, improvement ratings are preferable.

**Multiple Scales**

There are two cases which use multiple scales, Akam and Kelkoo 2. Akam uses four scales, asking the user to rate the item according to four criteria, “user-friendliness”, “price/gain”, “quality” and “overall”. A problem with this approach is that it enables the user to give contradictory ratings. For instance, look at the rating below. The overall rating (“Total”) does not reflect the detailed ratings above.
Kelkoo 2 uses seven different rating scales, partially adapted to the product categories. Of the seven rating scales, six use a 1 to 5 scale, and one uses a binary scale asking the user whether (s)he would recommend the product or not. Kelkoo 2 suffers from the same problem as Akam in that you can give contradictory ratings. Below is a screenshot of how monitors are rated (English translation to the right).

Ratings Across Cases
Due to the lack of number of ratings, there simply did not exist enough rated products, with a sufficient number of users having rated them, in order to create a trustworthy set of data. Hence, this step was skipped.

Rating Distribution
The table below shows how the ratings are distributed within the selection of items used in the chapter on shilling attacks. In order to get a better visual overview, the graph below the scale displays rating distribution as a graph. Due to the lack of number of rated items, DinPris is not included in the below calculations.
As both the table and graph shows, there is a clear tendency of ratings being high. Akam has a rather extreme distribution, 50.63 percent of the ratings being 4.5. Kelkoo has a bit more even distribution, but the majority of the ratings are still high. Komplett resembles the pattern of Akam, 40.48 percent of the ratings being 4.5. If you split the scale in half (or almost half, 1 to 3 and 3.5 to 5), you get some interesting numbers. Akam has 8.44 percent of its grades from 3 and down, and 91.56 percent from 3.5 and up. Kelkoo has 34.37 percent from 3 and down, 65.63 percent from 3.5 and up. Komplett has 7.14 percent from 3 and down, 92.86 percent from 3.5 and up. Clearly, distribution of ratings is very uneven, but with some variations between cases.
5.5 Confidence
This section will first present which confidence-building measures are used in order to explain why (or why not) an item is recommended, then take a closer look at how the average ratings are presented.

Confidence-Building Measures
All of the cases employ varying degrees of extra confidence building measures to enable the user to get further product information beyond the user-created ratings and reviews.

Akam
- Extensive technical product information.
- Picture gallery.
- Links to external product information sites, manufacturer’s site, expert reviews and the like.

DinPris
- Extensive technical product information.
- Option of a comparison of technical specification between two or more product of same type (i.e. two mobile phones).
- Links to external product information sites, expert reviews and the like.

Kelkoo
- Minimalist technical specifications.

Komplett
- Extensive technical product information.
- Links to manufacturer’s site and external expert reviews and tests.
- Option of a comparison of technical specifications between two or more products of the same type.

Presentation of Ratings
As my cases were being studied, I identified three different contexts in which the average rating is displayed. The first two contexts are whether the number of users rating the item is
displayed in addition to the average rating when (1) the item is in a list of other items and (2) when the item’s main information page is displayed. The last context is whether it is possible to view a list of all separate ratings for an item (each user’s ratings).

<table>
<thead>
<tr>
<th>Case</th>
<th>Page</th>
<th>Number of reviewers displayed when item is in a list with other items</th>
<th>Number of reviewers displayed on product information page for a single item</th>
<th>List of separate ratings for a single item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akam</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DinPris</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Kelkoo</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes/no</td>
<td></td>
</tr>
<tr>
<td>Komplett</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 8: Display of number of ratings.

The table shows that displaying the number of ratings is done in three out of four cases. Komplett does not display the number of ratings in any context. All cases except Kelkoo (when rating without writing a review) offer the user the opportunity of viewing all separate ratings done by all users on one specific item. When giving a rating in combination with a written review on Kelkoo, viewing separate ratings is possible.

### 5.6 Fading

This section describes the fading techniques used by the four cases.

**Akam**

Akam does not associate a date with any of its items. Any date-related operation (i.e. “New Products”) is hence not available. Fading is done by manually tagging the item as “outdated” when the product is no longer manufactured or has disappeared from the market. Products tagged as outdated are still available using the site’s advanced search option.

**DinPris**

DinPris creates its database of products based on product information from web shops with which they cooperate. A list of products is read from these web shops once a day. As the products are no longer on sale, they automatically disappear from the DinPris website.
**Kelkoo**

A request was sent to Kelkoo regarding this question, but an answer was never received. Deducing how fading is implemented by analyzing the web site itself is not possible. Hence, due to the lack of data, this issue remains unsolved.

**Komplett**

Since Komplett is a web shop, its products are naturally removed from the site when they no longer sell it. The site does offer a list of new products and the option of subscribing to an RSS feed with information about new products.

### 5.7 Written Reviews

This section discusses the results of the survey in which participants were asked to judge the usefulness of a set of written reviews.

All of the participants were between 22 and 28 years old. Of the 9 individuals participating, 66% were male and 33% were female. The self-reported general computer skills averaged 4.11 of which 1 was “very little knowledge” and 5 was “very good knowledge”. This average is a bit high, but on the other hand only individuals with a certain minimum of computer knowledge were selected in order to avoid a situation in which a participant did not know what (s)he was actually reviewing. On average the participants reported to have visited recommendation sites similar to the ones studied here 5.11 times per month, the lowest being none and the highest being 20.

A total of 33 reviews were judged by the participants. This is lower than the initial 36 reviews anticipated. The reason for this was the simple lack of written reviews at DinPris and Kelkoo. Each of the reviews was judged by 3 participants. The below table shows the average rating given by the participants on a scale from 1 to 5 where 1 was “very useless” and 5 was “very useful”.

<table>
<thead>
<tr>
<th>Case</th>
<th>Products reviewed</th>
<th>Reviews done</th>
<th>Average rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akam</td>
<td>9</td>
<td>27</td>
<td>2.85</td>
</tr>
<tr>
<td>DinPris</td>
<td>8</td>
<td>24</td>
<td>2.08</td>
</tr>
<tr>
<td>Kelkoo</td>
<td>7</td>
<td>21</td>
<td>1.76</td>
</tr>
<tr>
<td>Komplett</td>
<td>9</td>
<td>27</td>
<td>2.59</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>33</strong></td>
<td><strong>99</strong></td>
<td><strong>2.36</strong></td>
</tr>
</tbody>
</table>

*Table 9: Ratings of reviews.*
The results vary from case to case. Akam and Komplett hover a bit below 3 (the median of the rating scale), whereas DinPris is around 2 and Kelkoo on a very low score of 1.76. The average review rating of 2.36 should indicate that the participants found the reviews to be a bit below medium in usefulness. Taking variations between cases into consideration, the participants’ measure of usefulness of the reviews is a bit low.

5.8 New Weaknesses

Due to the fact that the collection of weaknesses gathered from social navigation theories and previous research presented in the theory section is rather extensive, I do not expect to add too much to this based on the limited number of websites reviewed. I have however discovered four weaknesses not described in the theory section.

Ads

The first weakness concerns the placement of ads on Kelkoo. Below is a screenshot of how the list of laptops is presented on the site.
The picture shows the upper part of the screen displayed when selecting the menu item “Laptop”. The list of laptops continues below the one seen at the bottom of the picture. The areas of focus are shaded. First, the upper row of products from DELL can potentially be mistaken for being part of the site’s menu system, thus prompting the user to unwillingly click on the ad. As opposed to Akam and DinPris, separation of ads and the rest of the pages’ content is not clear (Komplett does not use ads except for products they sell themselves).

Second, the list “Sortér etter…” (“sort by”) has three major laptop suppliers on top of the list, followed by an alphabetical overview of other companies (the list can be extended by clicking the link “Vis flere” (“show more”) below). This could be explained by these three companies being among the largest suppliers. On the other hand this arrangement easily gives the user the impression that these companies have paid to get their names on the top of the list, especially since the number of laptops from these companies (shown in parentheses after the name) should put HP Compaq on top and Acer second.
Distortion of Ratings and Price Comparison

A second weakness, which I must say I did not expect to find, is that not all items on the Kelkoo website can be rated. Below is a picture of a random item from the list of routers found on the web site.

![Picture 13: Item that cannot be rated at Kelkoo.](image)

This item cannot be rated, and there is very little information on the website distinguishing items which can and items which cannot be rated (the upper right link on picture 12 is not present on picture 13). According to Kelkoo, this is because they only keep a database with extensive product information (which includes the ratings) for some of their products, whereas for other products only a simple list of prices, product names and a short description given them from the web shops offering the products are kept. This distinction could be rather confusing to the user. In addition, when clicking on the linked name of an item that cannot be rated, you are taken directly to a web shop selling this product, whereas clicking on the linked name of an item that can be rated takes you to Kelkoo’s own information page on the item.

Barrier to User Entry

All of my cases suffer from lack of ratings. In the case of Komplett, rating is made more complicated by the fact that you need to be a registered customer in order to give a rating and write a review. Kelkoo uses a mixture of this, you do not have to register in order to rate a product, but you need a Yahoo! account in order to write reviews. Registering a Yahoo! account is cumbersome due to lack of user friendliness in the registration process. Second, you are forced to sign up for a new Yahoo! e-mail address, which has absolutely nothing to do with writing reviews of products. Akam and DinPris do not require any registration.

Contradictory Ratings

Akam and Kelkoo offer ratings on multiple scales. However, they have not put in place a mechanism which ensures the user doesn’t give contradictory ratings. For instance, it is possible give an item a low score on a set of specific properties but give the same item a high
overall score. This would most likely confuse other users, does the reviewer recommend the product or not?

5.9 New Solutions
My next step was to search for new solutions to the challenges described in the theory section. I have identified four solutions not yet discovered in the theory section. Some of the solutions found were surprisingly simple. This section will describe the new solutions.

Multi-Dimensional Rating
By studying the cases I’ve found that they do not only offer the option of rating the products themselves, but several other items within the CF system, thus creating a multi-dimensional rating environment. Below is an overview of items which can be rated in the cases which were examined.

- The products themselves.
- (In the case of Akam, DinPris and Kelkoo): These sites offer price comparison across different web shops. In addition, they give the users the option of rating the web shops themselves, thus expanding item information beyond the item itself but still with relevance to a potential customer.
- (In the case of Akam and Komplett): Akam offers the users the option of correctional user feedback (meta-rating). On Akam, for any item in their database, a user can give feedback on any of the information displayed about it, such as the technical information, other users’ reviews and so on. Komplett offers a somewhat similar service, by enabling users to rate other users’ reviews on a binary scale.

Availability and Cooperation Across Sites
Akam is a service available at several web sites. For instance, Hardware.no and PrisGuide.no displays the exact same product information, even the graphic is the same. This has the effect of exposing the product information across several websites, which should result in more users, which again should result in more ratings and reviews.

Manual reviewing of user-created content
Akam and DinPris use manual review of users’ reviews and ratings. Each user-created review is read and accepted/denied before it is displayed on the web pages.
**Distorted Text**

When rating an item or writing a review on Akam, distorted text is used. This protects against potential bot attacks, since bots have a very hard time reading the distorted text. Below is a screenshot of how this text looks.

![Distorted text at Akam.](image)

*Picture 14: Distorted text at Akam.*
6. Analysis and Discussion

People mistakenly assume that their thinking is done by their head; it is actually done by the heart which first dictates the conclusion, then commands the head to provide the reasoning that will defend it.

- A. de Mello

According to theory, social navigation can help turning a space into a place, but what kind of place? In the preceding chapters I have analyzed four cases in light of known social navigation weaknesses in order to assess how susceptible these cases are to these weaknesses. In this chapter I will discuss each known weakness and the newly discovered weaknesses and solutions. The goal will be to try to give an answer to the research question, how reliable are web-based recommender systems as an aid for making our choices?

6.1 Shilling Attacks

According to our theory, the fewer the number of items rated, the more susceptible this item is to a shilling attack. Additionally, the fewer ratings a rated item has, the more susceptible to a shilling attack. All my cases suffered from lack of rated items, the case with the highest number of ratings being Kelkoo with about half its products rated. The average number of ratings per rated item was not very high either, with Kelkoo again on top with 7.36. Even if a shilling attack was done manually, faking a rating on Akam, DinPris and Komplett would be rather easy (excluding other mechanisms for stopping a shilling attack). In the most “difficult” of these three, Komplett, if say ten people performed a shilling attack, they would account for 74.63 percent of all the ratings on the average item. In the case of Kelkoo, a manual shilling attack on the average item is a bit more difficult, one would require 7.36 fake ratings to account for half of the ratings on an item, but still, this number is not very high. The distribution of ratings shows that most items rated are rated by only a few users and few items are rated by many. We can conclude from this that all the cases, to a varying degree, are susceptible to shilling attacks when excluding other protective measures.
What implications does this have? First, due to the susceptibility to shilling attacks, when looking only at the number of ratings, the credibility of these rating systems is greatly reduced. With variations between cases, one should be cautious in trusting the recommendations given. Second, due to the massive lack of ratings found on Akam, DinPris and Komplett, the sites’ value as a recommendation system is greatly reduced. There is a basis for claiming that these sites barely can be regarded as a functional CF system.

Interestingly, in a study of the online book store Amazon, Hu, Pavlou and Zhang finds a pattern in the distribution of ratings similar to the one discovered in this thesis (Hu, Pavlou and Zhang, 2006, p. 325-326). Most items rated are rated by few users and few items are rated by many. The number of ratings was significantly higher than what was found in the cases presented here. However, this could point in the direction of a general trend within CF environments, namely that many items are rated by few and few items are rated by many.

On the other side, the cases studied do provide some protective measures to guard against shilling attacks, such as distorted text and manual review of ratings. While these measures do not help the sites gain more ratings, they act as a quality assurance of the ratings given. Hence, the sites are not sitting ducks for a shilling attack, and there is therefore a basis for claming that one can put trust in the ratings given. Additionally, and quite ironically, one could claim that a large-scale lack of ratings may actually to a degree protect against shilling attacks. Take for instance DinPris, which uses manual reviews of ratings. If one were to perform a shilling attack on this system, the sheer lack of ratings combined with the manual review would quickly cause the reviewers of the ratings to sound the alarm. Lack of ratings makes a CF system more susceptible to shilling attacks, but if the number of rating is sufficiently low, discovering a shilling attack becomes easier.

The lack of ratings found in the cases studied in this thesis, taking into consideration that only a selection of items was studied, points in the direction of the sites being in varying degree susceptible to shilling attacks. On the other hand, the sites do employ counter measures, helping the ratings gain some degree of credibility.
6.2 Barrier-to-entry
According to theory, CF systems are prone to the barrier-to-entry challenge. As shown in the preceding section, the cases studied suffer to varying degrees from this challenge. On the one end of the scale, DinPris has only .52 percent of the selection of items rated. On the other end, Kelkoo has about half of its selection of items rated.

The barrier-to-entry challenge in these cases is not that new items have a hard time gaining a rating, but rather that most items are not rated. The effect of this is straightforward. The CF system’s value as a recommendation system is reduced. Of the known techniques which can help items gain more ratings, very few were employed. To help solve their problems of lack of ratings, implementing these techniques might help the CF systems gain more user ratings. Interestingly, there is no correlation between those sites which implemented the highest number of the known alleviating techniques and the number of ratings the items on these sites received. There could be several explanations for this discrepancy. One explanation would be that the techniques do not have the desired effect, or that the techniques are implemented incorrectly or incomplete. An example of the latter can be said do be found in two of the cases; the usage of passive data collection is poorly utilized. Using hiding of newly added items would most likely not have a significant effect, since all items generally suffer from a lack of ratings. A different approach would be to look at effects outside social navigation theory. For instance, lack of visitors to the sites constitutes a viable explanation. In addition, design and usability affects the way visitors use the sites, which again could affect the number of users rating items. However, in order to gain a thorough understanding of these effects, further research is required.

All the cases suffer from the barrier-to-entry problem. Few techniques for alleviating this challenge are employed. However, since there is no correlation between the number of alleviating techniques used and the number of ratings received, one might look outside social navigation and the CF system itself in order to gain a further understanding.

6.3 Conformity
Existing research has shown that users, when rating an item, are affected by other users’ opinion of an item. The only identified method of alleviating this challenge is to give users the
option of hiding items’ ratings, which was not found in any of the cases. This leads to the conclusion that in these cases, influence by other users’ opinions may easily occur.

However, this effect might be reduced by the lack of ratings. If you’re the first to rate an item, you cannot be affected by other users’ opinions, since they do not exist. Looking at it from the bright side, lack of ratings may actually have a positive effect on the conformity challenge.

Finally, one aspect of conformity which might be further explored is whether users are more likely to rate an item with an existing rating or a high number of ratings than an item with no or very few ratings. If a CF system exposes items rated by several users to a higher degree than items rated by few users, one could have a basis for claiming this. On the other hand, my results on the distribution of ratings, which show that most items are rated by few users and few items are rated by many, points in the opposite direction. Work done in this thesis does not fully cover this topic; it was rather discovered through the work. If future research can prove than a high number of existing ratings causes a high number of new ratings, this could be a valuable basis for improving existing CF systems and extending social navigation theory.

None of the cases have any protection against the issue of conformity. However, through this work, I discovered the unexplored issue of whether users are more likely to rate an already rated item versus an item that has no ratings.

6.4 Rating Scales
The next step was to study the rating scales themselves. As social navigation does not offer sufficient theory around how these should be implemented, I had to move outside the realm of social navigation and make use of theory around surveys and questionnaires.

The results were mixed. In relation to theory, some faults were found, such as unequal intervals, unbalanced rating scales and assumption made by the question. On the other hand, there were no occurrences of forced choice and the number of points on scales was within what theory suggested. The most interesting finding was that two of the cases simply do not label their rating scales, and in one case, there is absolutely no text describing what the site expects of a review. This poses a bit of a problem when survey theory is applied, since these theories expect that there actually is a label. However, there should be basis for claming that
lack of labelling of scales could cause a higher degree of subjectivity in the ratings. One can assume that most people will understand that one star equals the lowest rating and five stars equals the highest rating, but how good is three stars? This lack of labelling of scales could bias the results since there is no common anchor point of what the different number of stars actually signifies.

The second implication I will associate with social navigation theory. In social navigation theory, there is little focus on how questions are asked and how rating scales are formed. The rating scale itself affects the results of a survey, and a biased rating scale could have biased results as the outcome (Friedman and T. Amoo, 1999). Since the vast majority of CF systems base their recommendations on data gathered from other users through the use of rating scales, this issue should be of great importance within the practical implementation of social navigation rating systems. However, in the author’s experience, little or no light is shed on this issue within social navigation theory.

The second experiment done on rating scales was to look at the distribution of ratings. With some variations between cases, the results showed that there is a clear tendency of most of the ratings being high within the selection of items studied. These results are consistent with studies of other rating systems. Hu, Pavlou and Zhang, in their study of ratings at Amazon, found the same pattern (Hu, Pavlou and Zhang, 2003, p. 326). Chevalier and Mayzlin, in a study of both Amazon and Barnes and Noble, also discovered this pattern (Chevalier and Mayzlin, 2003, p. 13).

Finding a cause for this distribution is very complicated, requiring a study in itself. Several aspects of CF systems have been outlined in this thesis, and a large part of them are candidates for possible explanations. The two studies referred to above do not attempt to explain this phenomenon either. However, and again moving outside social navigation theory, marketing research can give a partial explanation of why this phenomenon occurs. According to Godes and Mayzlin, very dissatisfactioned customers and very satisfied customers are most likely to engage in word-of-mouth (Godes and Mayzlin, 2004, p. 547). This however, should result in a U-shaped graph, whereas my results only show a majority of positive ratings. Hence, marketing science only explains half of the phenomenon. I therefore conclude that further research may be required in order to further understand this phenomenon.
A second important aspect is what implications the uneven distribution of ratings has. First, a good rating may not be as good as it seems. After all, most ratings are good. Hence, the trustworthiness of ratings is reduced. Hu, Pavlou and Zhang, through analyzing the distribution of ratings of products which have a mean score of the scale median (i.e. an average score of 2.5 on a 1 to 5 scale), show that this average does not reveal a product’s true quality, since the ratings do not concentrate on the mean (Hu, Pavlou and Zhang, 2006, p. 327). To further explain, they show that a median score is more likely a result of a large number of very low ratings and a few higher ratings, rendering the median rating unrepresentative for the product’s “true” quality but rather reflects the balance point of very different opinions. Unfortunately, due to the simple lack of ratings I was unable to perform such an analysis on my cases, the number of items with an average median rating was simply too few to make the results statistically viable. On the other hand, taking into consideration that theirs and my results match in rating distribution, there is a small basis for claiming that the “true” quality of a product is not necessarily present in my cases either.

In my research into the cases rating scales, I have made two important discoveries. First, that social navigation theory does not give an explanation or emphasise the importance of the rating scales themselves. Second, ratings both in my cases and in other studies tend to be unevenly distributed towards most ratings being high.

6.5 Confidence
The next issue is that of confidence. According to what I identified in the theory section, simply displaying ratings is not enough in order for the user to gain confidence in the system. In my research I identified two different ways in which confidence is gained, namely information about an item beyond users’ ratings and reviews and how ratings are presented.

As shown, most of the cases use links to technical information, external product information sites and external expert reviews. This helps both to build confidence in a recommendation and enables the user to browse further information about the product. Interestingly, some of this information (such as technical specifications) is strictly speaking outside the realm of social navigation. One can conclude that a combination of social navigation techniques and non-social navigation information is an absolutely useful symbiosis for increasing the
credibility of a recommendation and enabling the user to gain further insight into the item in question.

When it comes to displaying the inner logic of recommendation systems, I chose to look at whether my cases display how many users had rated a specific item. Since neither of my cases use targeted recommendations, I could not see how further explanation of inner logic could be done. With some variations, the cases did fairly OK. None of the cases did consistently display number of users rating an item whenever an average rating was displayed, but three of the cases did whenever the main page for information about a single product was displayed. All in all, there is a small room for improvement, but none which will be a radical change for the better. This is with the exception of Komplett, which at no point displays how many users have rated an item.

All in all, my cases do pretty well in gaining confidence in the recommendation through extra non-social navigation information and in how average ratings are displayed. The combination of social navigation information and non-social navigation information could be very beneficial for the user.

### 6.6 Fading

The issue of fading of items is well recognised in social navigation literature. My cases use two different approaches when an item is removed due to old age. They are either manually removed or automatically removed based on a third-party product database.

The question arises whether this can be considered fading at all. Let’s first introduce into the discussion the fact that two of the cases use increased exposure of new items through a “New” list. If an item’s life goes from such a list, moving down the list, appearing within the regular menu structure in its appropriate category and finally being removed because of old age, it would match the word “fading”. Hence, two of my cases implement fading. The remaining two use a socialist binary approach, either you’re in there on an equal level with everybody else, or you are removed.

On the other hand, one could look at fading as only using the latest ratings on an item when calculating the average rating, i.e. only base the average rating on the last ten user ratings
(none of the cases does this). This would enable an item to more easily change its average over time. This aspect of fading poses a problem when rating computer equipment. As long as product itself does not change over time, why then should one only record the latest ratings? Compare this, for instance, to a TV series, which may change over time. In this case, recording only the latest ratings would make sense in order to more correctly reflect the TV series as it changes. One therefore has a basis for claiming that this type of fading should only be implemented with great caution on items which themselves do not change over time.

Two of my cases implement fading in the item’s journey from introduction to the site and till it is finally removed. The other two do not implement fading. None of the cases use fading in the sense of utilizing only the latest ratings when calculating an average, but when rating computer equipment, this might be a wise choice.

6.7 Written Reviews
According to the survey of how useful a group of individuals found the written reviews to be, with some variations, there is a general lack of usefulness of the written reviews in my cases. In addition, two of the cases suffered from a severe lack of written reviews. Akam and Komplett require the user to write a review in combination with rating on a scale, and hence the number of ratings and number of reviews are equal. DinPris and Kelkoo on the other hand, do not require these two operations to be done in tandem, which is reflected in a severe lack of written reviews. Among the selection of products used in this work, there were not enough written reviews on DinPris and Akam to reach the goal of nine reviews wanted for the survey.

In the introduction I talked about the concept of community wisdom. According to my results, the community (the written reviews) idea can be contested. Since the survey shows that there is a general lack of usefulness in the written reviews, the community may not be that wise. There is however a second possible explanation. It does not contradict the idea that the community may not be that wise, but offers additional insights. What is useful to me may not be useful to you and vice versa. Hence, one user may find a written review highly useful, whereas another user may find the same review less useful. This creates a challenging dilemma, namely that the same written review may be both useful and useless depending on subjective factors in the reader. The most likely explanation for my results is a combination of
these two factors. The author’s own reading of the reviews presented to the participants and the fact that some of the reviews got horribly low ratings would support the theory of lack of wisdom in the community. However, subjectivity cannot easily be explained away and should be considered a factor in the results.

The combination of lack of written reviews and the low degree of usefulness found has important implications on the CF systems studied. First, it leaves the systems open to shilling attacks in the form of fake written reviews. Second, the systems’ value as recommendation systems is reduced since there is a general lack of reviews, and when a review is found, the survey shows that there is a high probability the review lacks usefulness from the reader’s perspective. Third, there is a lack of confidence. The users’ confidence in the recommendation systems would most likely be reduced when they found a low degree of usefulness in the recommendations given.

The lack of usefulness found in the reviews can be explained both by lack of community wisdom and subjective factors. The general lack of written reviews makes the CF system more vulnerable to shilling attacks. The combination of low usefulness in reviews and lack of reviews causes the CF systems to be less valuable and decreases confidence.

### 6.8 New Problems

**Ads**

Kelkoo has a blurry distinction between ads and the site’s menu structure. Even though this is not a direct part of the CF system itself, the blurry distinction could have an effect on the users’ confidence in the site and hence the confidence in the CF system’s recommendations. If the user cannot trust the site being impartial when it comes to the various manufacturers making the products rated within the CF system, the user may also start questioning whether the ratings themselves are influenced by this impartiality.

On the other hand one could argue that a site should be allowed to place ads and that the non-alphabetical list of manufacturers is due to some reason I did not discover. However, it does not matter whether there is a reasonable explanation for this arrangement as long as the user gets the impression of the site menu structure being sponsored by the companies in question.
It is the users’ perception, not the site’s intentions, which eventually decides whether the users put their trust in the site or not. We cannot expect a recommendation system on the web not to use ads. However the creators of an online CF system should take care to portray an image of impartiality. In the case of Akam and DinPris (which both also have ads on their sites), there is a clear distinction between ads and the content produced by the CF system, which gives the user a clearer impression of impartiality. As discovered earlier, this is yet another factor, strictly speaking, outside the CF system itself, but which still has an indirect effect on the system and hence should be taken into account.

**Distortion of Ratings and Price Comparison**
The rather odd solution of mixing ratings and price comparison on Kelkoo has two effects. First, it disrupts usability when the user (including the author) has trouble distinguishing a recommendation from a price comparison. Since this type of usability is not a topic in this work, we’ll leave the discussion there. Second, it could have the effect of reducing confidence since the user becomes uncertain about what is a recommendation and what is a simple price comparison. One could argue that this mix is a foolish design choice which affects the confidence of the recommendation system. Both Akam and DinPris offer price comparison, but have a clear distinction between recommendations and prices.

**Barrier to User Entry**
Requiring a user to register on a web site before starting to use it is not uncommon. In the case of Komplett, which is a web shop, requiring registration before use is natural. However, Kelkoo is overdoing it by requiring the users to create a new e-mail account before being able to post written reviews. One could argue that this is reflected in the large lack of written reviews found on this site. Unnecessary registration constitutes a barrier-to-user-entry. If the registration process is too complicated, fewer users will register and fewer reviews will be written. Hence, this barrier-to-user-entry causes the system to be more vulnerable to shilling attacks. As with the ads above, barrier-to-user-entry is not directly a part of the CF system itself, but could have a great effect on how many users actually write a review or rate an item in the system and hence becomes important in our context.

**Contradictory Ratings**
According to our theory, rating on multiple scales can give the user a valuable broader view of an item. However, when there is no mechanism to prevent contradictory ratings when using
multiple scales, the user could potentially lose confidence in the rating. How can you trust a rating which logically does not make sense to you? Solving this problem can easily be done on a technical level. Whenever multiple scales are used, designers should ensure that contradictory ratings are not possible.

6.9 New Solutions

Multi-dimensional rating
Several of the cases offer rating across multiple dimensions. These ratings can be divided into two categories.

The first category is rating items outside the products themselves, such as the web shops where these products are sold. These ratings help potential buyers get an impression of other users’ experiences with the web shops, and since reliance on the seller is important in a sales situation, these ratings feed the potential buyer with valuable information. According to social navigation theory, confidence is of great importance, and other users’ experiences with a web shop can help other users gain confidence (or lack of confidence) in a potential seller. Hence, this combination of ratings should be very valuable.

The second category is meta-ratings. According to the principle in social navigation theory of CF systems being self-correcting, this should have a good effect. For instance, providing feedback on other users’ reviews could help reduce the credibility of reviews which are not useful to other users, contain nonsense or the like. On the other hand, this sort of meta-rating runs into a fighting-fire-with-fire dilemma. As shown in this thesis, CF systems both in theory and practice come with several potential faults. Hence, by adding a second layer of rating one might not achieve a better CF system at all, rather a second set of problems. It is hard to conclude one way or the other, but a greater awareness of this dilemma would in any case be useful.

Multi-dimensional ratings could be a very useful tool for enhancing the value of the CF system. On the other hand, meta-ratings should be used with some caution, and could potentially simply create a second layer of problems rather than resolve the existing challenges.
Availability and Cooperation Across Sites
Availability across sites can help a CF system gain more users. In other words, using the same CF system, with its underlying data collection mechanisms and databases, across several web pages, could cause an increase in the number of users, which again could cause more ratings which finally could help alleviate the barrier-to-entry challenge. This concept can be taken a step further, namely cooperation across separate product recommendation systems (none of the cases actually do this). To illustrate, let’s say Akam and Kelkoo decided to cooperate and share their ratings and reviews. If a user rated and wrote a review of an item on Akam, this rating and review would also appear on Kelkoo. Thus, more reviews and ratings will appear on both sites, and this would again help alleviate the barrier-to-entry challenge. However, these two techniques are simply methods of getting more users to a site, and there are thousands of ways in which increasing site traffic can be done. On the other hand it illustrates an important point in social navigation theory, namely that solutions to a challenge (in this case the barrier-to-entry problem) can be found outside social navigation itself.

Manual reviewing of user-created content
Some of the cases use manual reviewing of user-created content. This method touches upon two of our challenges, namely shilling attacks and confidence. Manually reviewing user-created content can guard against shilling attacks, in the case when the person reviewing notes a pattern which raises concern of a shilling attack, i.e. many reviews from the same person, several positive/negative reviews for a single item. Additionally, manual reviews can help ensure quality in the written reviews and hence be a confidence-building measure. However, this method is naturally very time consuming and requires a certain number of staff and hence increased costs.

Distorted Text
The use of distorted text on the Web has become widespread. Since distorted text can guard against bot attacks, it could be a very valuable tool to protect against shilling attacks, and hence increase the general confidence in the CF system itself.
7. Conclusion

*Science is always wrong. It never solves a problem without creating ten more.*

- G. B. Shaw

The goal of this work was to try to give an answer to the question of how reliable web-based recommender systems are for making our choices. The approach was to first collect known weaknesses of social navigation rating techniques, analyze four Norwegian recommendation web sites in relation to these weaknesses, and in addition survey the web sites for any new strengths or weaknesses not yet discovered in the theory section. Pulling all the strings together is not an easy task, and this work might pose more questions than it answers. In this concluding chapter, I will start by giving a very short overview of the results, explain why these can be extended beyond the cases studied, present what implications these results have for users, social navigation theory and designers and finally suggest some topics for future research.

With some variations between the cases studied, the results can be summed up as follows. Eight known weaknesses of social navigation rating systems were identified, of which two were discarded due to lack of relevance in this thesis. In addition, a method for surveying the usefulness of written reviews was created, based on known weaknesses of written reviews and similar work presented in other papers. Of the six known weaknesses identified in the theory section, all the cases are prone to five of them. All cases suffer from the barrier-to-entry challenge, and are hence open to shilling attacks. None of the sites offer any protection to the challenge of conformity. A study of the rating scales shows an uneven distribution of ratings. Finally, fading is poorly implemented or missing. Only a few of the identified counter measures to these challenges were implemented. On the bright side, all the cases do well on the issue of confidence, since they all offer the users product information outside the CF system itself. Moving on to the survey on written reviews, there seems to be a general lack of usefulness of these reviews, keeping in mind some degree of variation between cases. Finally, a set of new strengths and weaknesses was discovered. I found four new weaknesses. Poor placement of ads and distortion of price comparison and recommendations causes reduced
confidences. An unnecessarily complicated registration process increases the barrier-to-entry problem. Finally, the possibility of giving contradictory ratings affects the confidence of the rating scales. Four new solutions were found. Multi-dimensional ratings could increase the usefulness of the rating scales. Cooperation across sites could result in more users and hence help alleviate the barrier-to-entry challenge. Manual control of ratings and written reviews helps increase confidence. Finally, the use of distorted text can protect against shilling attacks.

A very important question is whether these results can be applied to other CF systems. The answer is twofold. First, these specific results can of course only be used when judging the four cases in question. On the other hand, by highlighting the weaknesses of social navigation recommender systems, identifying these in other cases becomes easier, and when a weakness is identified it can be corrected. Being aware of the pitfalls will help you avoid them and hence these results can be very valuable in the studying, implementation, correcting and use of other CF systems.

7.1 Implications for Users
What implications do the results of this study have for users of online recommendation systems? We can split this question into two parts, implications for the users of the cases studied and implications for the general use of online recommendation systems.

With some variations, the cases studied in this work all do rather poorly as recommendation systems. In fact, it greatly surprised the author. I have taken a broad approach looking at a wide range of different aspects of the social navigation techniques employed in the four cases, but weaknesses are many and strengths are few. To directly answer the question of how reliable these four cases are as aids for making our choices, I will conclude that they are highly unreliable. This is of course limited to the product categories studied in this work, and cannot necessarily be extended to all product categories found in these CF systems. The implications for the users of the four cases are naturally that they should be very sceptical when basing their choices on the recommendations given. Swearingen and Sinha writes that […] users preferred recommendations made by their friends to those made by online systems (Swearingen and Sinha, 2002). In these four cases, the results presented in this study support this preference.
What about the general use of social navigation recommendation systems? The results presented in this thesis can naturally only speak for the four cases studied. On the other hand, taking into consideration the unreliability of these four systems, one can deduce that it is likely that there are other online recommendations systems which exhibit, in varying degrees, the weaknesses presented in this paper. On the other side, there is nothing in the results presented in this paper arguing against the existence of robust and well-functioning recommendation systems. However, it would be highly advisable to users of online recommendations systems, whose reliability is still unknown, to be very cautious in trusting the recommendations given.

### 7.2 Implications for Social Navigation Theory

What implications do these results have for social navigation theory? Based on the results in this paper, I will take a closer look at space and place and the notion of CF systems being self-correcting. Finally, I will suggest two new approaches to the study of social navigation.

According to theory, social navigation can help turn a space into a place. But what kind of place? If the cases studied in this paper can be said to be a place, it would be a desert with the occasional oasis. The cases studied in this paper, keeping in mind that only a selection of products was surveyed, can only to a limited degree be said to have succeeded in transforming a space into a place. Additionally, according to social navigation theory, transforming a space into a place can cause a sense of ownership through the users’ contribution, which again could make users stay longer on the site and visit more frequently. However, deducing from the lack of ratings and reviews on these sites, there seems to be a lack of participating users. A second important aspect of CF systems is the notion of these systems being self-correcting. However, due to the general lack of rated items, the cases studied cannot be said to be self-correcting. If I were to place a fake rating of any of the items within the categories studied in this paper, statistically, the general lack of ratings would not cause it to be corrected.

The important impact on social navigation theory on the issues of space and place and self-correcting is that these two mechanisms do not work properly in the cases studied. The next natural question to ask is of course why? Even though it may disappoint the reader, I will not attempt to answer this question. The reason is simply that it is beyond this study, and trying to
answer it would be highly speculative. However, looking further into why social navigation techniques do not necessarily succeed becomes an interesting candidate for future research.

An interesting thought is that if there had been more users on these sites, a large portion of the challenges could have been solved. If more users caused more ratings, the barrier-to-entry problem could decrease and the CF system would be less vulnerable to shilling attacks. This leads us onto the path of a new way of studying social navigation systems. Most of the literature read by the author on social navigation focuses solely on the social navigation system itself. However, as shown in this paper, several factors which affect the social navigation system are found outside the system itself, such as for instance the number of users visiting the site. By looking at the social navigation system in context, one can separate three different parts, the social navigation techniques, the user and the technical environment. The social navigation techniques are the known social navigation methods themselves. The user is whoever uses the system. The technical environment is the conditions in which the social navigation techniques reside and through which the user interacts with the social navigation system. Graphical layout and user friendliness are examples of technical environment. All these factors influence each other. For instance, lack of user friendliness can cause fewer visitors, which again causes few ratings which results in vulnerability to shilling attacks. Hence, studying only one of these three factors will not give a complete picture. The author would therefore suggest a slight shift in how social navigation systems are studied. More light should be shed on the user and the technical environment in relation to social navigation systems, and how these three factors affect each other.

Whereas the first approach focuses on what should be studied, the second approach focuses on how. As illustrated by the usage of survey theory in the study of rating scales in this paper, social navigation theory itself does not fully encompass the study of CF systems. The usage of non-social navigation theory for the study of social navigation can be taken a step further. As already mentioned, social navigation theory relies heavily on the physical-world metaphor. In addition to this, social navigation is strictly speaking nothing new, it can be said to have existed as long as humans have. Thus, social navigation in the digital world can be said to be a new arena serving an age-old phenomenon, the new arena, however, having new properties achieved through technological advances. From this point of view, social navigation becomes the tool through which the new digital arena is created. However, studying this new arena is far beyond what social navigation theory can ever do alone. In fact, studying the social
navigation digital arena could include myriad sciences, taking much the same approach as within HCI, including any science which might contribute to a better understanding of the phenomenon at hand. Applying new sciences to the study of digital social navigation arenas could be very valuable in gaining a deeper understanding into the mechanisms at work in these systems and hence help improve social navigation theory itself.

7.3 Implications for Designers
What implications do the results of this paper have for designers? We can identify two important aspects, namely that implementing social navigation rating techniques alone is not enough and that designers should know what they are doing when creating a CF system.

Simply implementing social navigation techniques is no silver bullet for creating a successful web site. As described in the initial theory section, social navigation techniques have properties which indeed may help build a better web site, but implementing these techniques alone is not enough. Even though prone to several weaknesses, all of my cases do actually use social navigation techniques, but they still do not succeed properly as CF systems. In order to build a web site which successfully utilizes social navigation techniques, one should take into consideration factors outside the social navigation system itself. In line with implications for social navigation theory above, users and technical environment should be taken into consideration in order to create a complete picture of how the web site will function.

As seen in this study, some of the known weaknesses of social navigation recommender systems have known solutions, but several of these are not implemented in the cases studied. This leads us to the second implications for designers, namely that they should know what they’re doing when implementing social navigation techniques. The Web is a bit of a jungle when it comes to design, it is rather simple for an individual with some basic knowledge of web design to create a web site. However, when implementing social navigation techniques, the designer should have some knowledge about these techniques as well in order to avoid the known pitfalls found in social navigation.

7.4 Future Work
This section will present ideas for future work based on the results found in this paper.
Based on the first of the new approaches to social navigation suggested in the concluding sections above, further work could be done to understand the relationships and interactions between the social navigation techniques, the user and the technological environment. Through further knowledge into these interactions, valuable new insights into social navigation itself and the social navigation system in context can be achieved.

A second interesting area concerns the inclusion of other disciplines into the study of social navigation. Some already exist and have been used in this paper, such as Dellarocas (2003) (management science) and Godes and Mayzlin (2004) (marketing science), though none of these studies explicitly mention social navigation. There are two approaches through which this interdisciplinary work could be done. One is to look at existing research applicable to social navigation, such as is done in this paper. A different approach would be to apply a new discipline when doing new research on social navigation. In the latter case, using sociology in order to gain further understanding into how people use social navigation systems may be of special interest.

A third area which could be further explored is to do a comparative analysis of two social navigation sites, one functioning well and one with less success, in order to highlight which factors are important in order to create a successful social navigation web site. In the conclusion above, it is stated that some of the social navigation techniques employed in the cases studied did not work as intended, but I stop short of explaining why. By doing a comparison of two sites, this gap might be filled.

In the conclusion, one of the implications for designers was that they should have more insight into how social navigation systems should be implemented. Based on this thought, creating a set of guidelines for building social navigation systems and criteria for heuristic evaluation of existing systems could be very helpful. This paper may be a basis for creating such guidelines. However, further work is needed in order to create a trustworthy set of evaluation criteria and guidelines.

As seen in this paper, the survey on written reviews showed that the participants found these reviews not to be very useful. I do not attempt to explain why they were not found to be useful. Further work could be done in order to fill this gap. By understanding which aspects of written reviews users find useful and less useful, CF systems could be modified in order to
guide the authors of written reviews to write recommendations which are more useful to the readers.
References


Chalmers, M. 2003, “Informatics, Architecture and Language” in Höök K., D. Benyon and


Rose, K., (01.05.2007), *Digg This: 09-f9-11-02-9d-74-e3-5b-d8-41-56-c5-63-56-88-c0*, [online], Digg. Available at: <http://blog.digg.com/?p=74> [21.07.2007]


Wired, (2002.03.07), *Amazon Pulls Plug On 'Advice'*, [online], Wired. Available at:

Cases
Akam, <produkt.akam.no> [2007.06.17 – 2007.10.10].
DinPris, <www.dinpris.no> [2007.06.17 – 2007.10.10].
List of Illustrations

Pictures

Picture 1: Rating an item at www.break.com. .................................................................29
Picture 2: Recommending at www.amazon.com. ..........................................................30
Picture 3: Rating users at www.qxl.no. ........................................................................30
Picture 4: Structure of a CF system.............................................................................31
Picture 5: Presentation of a product on the Akam web site.........................................63
Picture 6: Presentation of a product on the DinPris web site.......................................64
Picture 7: Presentation of a product on the Kelkoo web site.......................................65
Picture 8: Presentation of a product on the Komplett web site....................................67
Picture 9: Rating scales................................................................................................76
Picture 10: Contradictory ratings at Akam..................................................................79
Picture 11: Multiple rating scales at Kelkoo.................................................................79
Picture 12: Ads closely knit into page content on Kelkoo............................................85
Picture 13: Item that cannot be rated at Kelkoo.............................................................86
Picture 14: Distorted text at Akam................................................................................88

Tables

Table 1: Rated versus not rated items...........................................................................69
Table 2: Average number of ratings per rated item.......................................................70
Table 3: Technical prevention of shilling attacks.........................................................73
Table 4: Employment of barrier-to-entry countermeasures..........................................74
Table 5: Overview of rating scales..............................................................................75
Table 6: Rating questions..............................................................................................77
Table 7: Rating distribution in percent........................................................................80
Table 8: Display of number of ratings..........................................................................82
Table 9: Ratings of reviews..........................................................................................83

Graphs

Graph 1: Rating distribution at Akam..........................................................................71
Graph 2: Rating distribution at Kelkoo.........................................................................71
Graph 3: Rating distribution at Komplett.....................................................................72
Graph 4: Combined rating distribution........................................................................72
Graph 5: Rating distribution..........................................................................................80
Appendices

Appendix 1: Survey of written reviews

The below survey was used when analyzing written reviews. In total, three different surveys were used, but they differed only in which products they linked to. Due to this similarity, only one of the surveys is displayed. The survey was done on the web.

<table>
<thead>
<tr>
<th>Spørreundersøkelse, produktanbefalinger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bakgrunnen for denne spørreundersøkelsen er å nærmere undersøke fire ulike nettsider for produktanbefalinger. På disse nettsidene finner du en oversikt over ulike produkter, samt andre brukeres erfaringer med disse produktene. Denne undersøkelsen er sentrert rundt skriftlige produktanmeldelser fra disse nettsidene.</td>
</tr>
<tr>
<td>Du vil bli presentert 11 ulike produktanmeldelser. Disse anmeldelsene er skrevet av normale brukere av nettsidene. Formålet med denne undersøkelsen er å vurdere hvordan du oppfatter disse produktanmeldelsene.</td>
</tr>
<tr>
<td>Produktene som presenteres er utelukkende datautstyr. Du vil lese produktanmeldelsene direkte fra nettsidene de befinner seg på. Ved å trykke på lenken til produktet, vil anmeldelsen åpnes i et nytt vindu.</td>
</tr>
<tr>
<td>Undersøkelsen starter med noen enkle spørsmål om alder, kjønn og datakunnskaper. Så følger vurdering av produktanmeldelser og til slutt et spørsmål om lignende sider du har besøkt tidligere.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alder og kjønn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min alder (år):</td>
</tr>
<tr>
<td>Kjønn: □ Mann □ Kvinne</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datakunnskaper</th>
</tr>
</thead>
<tbody>
<tr>
<td>På en skala fra 1 til 5, der 1 representerer &quot;svært lite kjennskap til&quot; og 5 representerer &quot;svært god kjennskap til&quot;, hvor ville du plassere dine egne datakunnskaper?</td>
</tr>
<tr>
<td>Svært lite kjennskap til □ □ □ □ □ Svært god kjennskap til</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vurdere produktanmeldelser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenk deg at du vurderer å kjøpe det produktet som presenteres. Hvordan vil du vurdere produktanmeldelsen du leser? Svaret gis på en skala fra 1 til 5, hvor 1 representerer &quot;veldig unyttig for meg&quot; og 5 representerer &quot;veldig nyttig for meg&quot;. (Trykk på den blå teksten for å få opp produktanmeldelsen i et eget vindu.)</td>
</tr>
</tbody>
</table>

**Produkt 1/11 (PC-skjerm)**
- Veldig unyttig for meg □ □ □ □ □ Veldig nyttig for meg

**Produkt 2/11 (Skriver)**
- Veldig unyttig for meg □ □ □ □ □ Veldig nyttig for meg

**Produkt 3/11 (Bærbar PC)**
- Veldig unyttig for meg □ □ □ □ □ Veldig nyttig for meg

**Produkt 4/11 (PC-Skjerm) NB! Trykk tekst merket 'In nlegg' rett til høyre for raden av stjerner**
- Veldig unyttig for meg □ □ □ □ □ Veldig nyttig for meg

**Produkt 5/11 (Bærbar PC) NB! Trykk tekst merket 'In nlegg' rett til høyre for raden av stjerner**
- Veldig unyttig for meg □ □ □ □ □ Veldig nyttig for meg
Produkt 6/11 (PC-skjerm)
Veldig unyttig for meg Veldig nyttig for meg

Produkt 7/11 (Skriver)
Veldig unyttig for meg Veldig nyttig for meg

Produkt 8/11 (Bærbar PC)
Veldig unyttig for meg Veldig nyttig for meg

Produkt 9/11 (PC-skjerm) Nede på siden under "Produktanmeldelser", med overskrift "Beste skjerm på markedet?"
Veldig unyttig for meg Veldig nyttig for meg

Produkt 10/11 (Bærbar PC) Nede på siden under "Produktanmeldelser", med overskrift "Gjør Jobben sin!"
Veldig unyttig for meg Veldig nyttig for meg

Produkt 11/11 (Stasjonær PC) Nede på siden under "Produktanmeldelser", med overskrift "Artig, velutstyrt liten sak, men...."
Veldig unyttig for meg Veldig nyttig for meg

Lignende sider
Hvor mange ganger har du besøkt sider lignende de ovenfor, der brukerne av sidene selv kan vurdere produkter, tjenester e.l.? Hvis du aldri har besøkt slike nettsider, la feltet nedenfor stå tomt.

Jeg besøker slike sider omtrent [ ] ganger per måned.

Ferdig!