Chapter 7

Conclusion

This thesis is grounded on an important observation: there is a disparity between how collaborative filtering is reseached and how it is deployed. The majority of research treats the scenario as a *static* problem: given a dataset, the quality of a particular algorithm's recommendations (measured as accuracy or precision) can be evaluated by training and testing the algorithm with partitions of all the data. Deployed recommender systems, instead, have to cope with a continuous influx of ratings, users, and content. The underlying data changes in size, sparsity, and may even become distributed differently; changes occur that affect performance and can neither be reproduced nor examined under static conditions.

Once the assumption of a static context has been removed, the methodology used to investigate CF needs to be redefined. In Chapter 4, we introduced a novel means of doing so, based on partitioning the data according to rating timestamps and simulating a deployed system by iteratively retraining CF algorithms with incrementally larger portions of the data. There are then a number of novel directions and uncovered results that can be examined when researching collaborative filtering. We have focused on three aspects: recommender systems' temporal accuracy, diversity, and robustness. Each aspect is highly significant: while accuracy has been the focal point of CF evaluation (and the primary tool for comparing algorithms), temporal experiments show that the way accuracy varies with time undermines the usefulness of work comparing algorithms solely on these grounds. Temporal diversity could not be explored from a static perspective, yet (especially in the case where it is missing) elicits passionate responses from surveyed users. Lastly, we determined that learning temporal behaviour and monitoring it for anomalies not only wards off a number of recommender system attacks, but forces attackers to select strategies that are both more costly (in terms of time taken to execute the attack) and less efficient. In the following section, we summarise the contributions we have made.

7.1 Thesis Contributions

At the broadest level, the contributions of this thesis fall into one of three categories:

- **Analysis.** We have shown, through extensive analysis of real user ratings, how CF data changes, including how summary statistics, similarity, and user behaviour fluctuate over time. While different datasets grow at varying rates, they all grow: observing these changes strongly motivates research into how the system *as a whole* performs over time.

- **Methodology.** We have designed a novel methodology for performing temporal collaborative filtering experiments. This method relies on partitioning the data according to the ratings' timestamps and incrementally growing the size of the training set.

- **Algorithms.** We have designed and evaluated hybrid CF algorithms that (a) increase the temporal accuracy, (b) bolster temporal diversity, and (c) secure recommender systems from temporal attacks.

The contributions of this thesis are relevant to (a) *researchers*, who are now challenged to build algorithms that stand the test of time (as well as those imposed by traditional evaluation metrics), and (b) *practicioners*, who may wish to augment their systems with features of this work, by, for example, overlaying a re-ranking algorithm on their CF prediction method. Recent work by Burke [Bur10] furthers the call for dynamic, temporal evaluations of CF by proposing a methodology that is similar, yet finer-grained, than the one we used throughout this thesis.

A general theme emerges from the algorithmic proposals we have made: whether we were focusing on improving accuracy, diversity, or robustness, our solutions proposed to treat users differently from one another. For example, the user-based switching algorithm (Chapter 4) improved overall accuracy by trying to improve each user's accuracy independently of the others; a similar solution was adopted to improve temporal diversity. When it came to defending a recommender system (Chapter 6), part of our proposal was monitoring users and comparing how they behaved with respect to the rest of the community, in order to identify misbehaving sybils. The centrality of users in our proposals reflects the variety of roles that users adopt when interacting with a recommender system: while some users are purely consuming content with the goal of obtaining better recommendations for themselves, other users are actually driven by a desire to help others' recommendations [HKTR04]. The key insight here is that there is a difference between the various system users; they are not all the same. CF research, on the other hand, has ignored this insight and designed "one-size fits all" solutions. In this thesis, we have departed from this approach by testing algorithms that vary how they compute predictions for different users.

The work we have done here is inherently limited by the data that we have used. Recommender systems may span a variety of different domains (both on and off the web and for a wide range of different types of items); however, our datasets only reflect online movie rental web sites. Our work has therefore focused on scenarios where users *explicitly* rate content (we do not use any implicit data). The assumption we hold is that these datasets are sufficiently *representative* of large scale recommender systems, and conclusions that we draw when analysing them are similarly applicable to other recommender system domains.

7.2 Future Work

As we saw in Table 2.1 (Chapter 2), the research problems relating to recommender systems are not limited to those we have addressed in this thesis. In this section, we discuss opportunities for future research. We divide them into two categories: the direct consequences of the methodology we have used

throughout this thesis and broader considerations on state-of-the-art CF research.

7.2.1 Using a Temporal Methodology

In this thesis, we focused on accuracy, diversity and robustness from a temporal perspective. The *approach* that we have defined, however, also offers a novel perspective on many other research challenges: attempting to solve them using a temporal methodology is likely to offer insights that were previously unavailable. Examples include:

The Cold-Start Problem. We explored the effect that highly connected users have on predictive accuracy and coverage and validated that only using them still achieves comparable accuracy results (Chapter 3). Cold-start users, who have no profile, may therefore be given a neighbourhood of these users until their profiles are sufficiently large to compute a similarity-based neighbourhood. Can these highly connected users be identified as they rate? To what extent do they vary over time and what effect does any variation have on system performance?

Serendipity. Being able to identify users who consistently seek out and rate new content may help finding the sources of *serendipitous* information. On the other hand, serendipitous ratings may be more prevalent in the sparser profiles. Herein lies a two-fold research problem: first, how can serendipity be measured? Second, is it possible to identify those who are the source of new ratings, trends, and who first rate what will later become popular content? Richer datasets may also offer finer-grained insights. In particular, recent work on multidimensional recommender systems may show why power-users emerge, and how they can best be used [ASST05].

Scalability. The mainstream approaches used to tackle the large number of users involves dimensionality reduction techniques [BK07]. Temporal patterns in neighbourhoods, however, can be taken advantage of to reduce the complexity of recomputing the similarity between every user pair. Identifying the active users in a particular moment can potentially be used to reduce the time complexity of computing recommendations. Furthermore, as discussed in Chapter 3, it is often the case that large proportions of the dataset are not used to generate predictions at all. Identifying and requiring only a small set of power users to generate accurate predictions would vastly reduce the scalability issues that recommender systems face [ALP+09].

Combining Multiple Goals. CF research has traditionally placed a high value on accuracy; in this thesis we designed mechanisms to augment accuracy over time. However, we also noted that both diversity and robustness are equally important. While our proposal regarding temporal monitors to secure recommender system robustness does not interfere with any underlying prediction or ranking algorithm, the diversity and accuracy algorithms may conflict with one another. Future work thus calls for designing and evaluating CF algorithms that meet a variety of requirements; for example, that they produce recommendations that are both (temporally) accurate and diverse. A simple approach to this particular example may entail using the switching algorithm (Chapter 4) to improve accuracy, while re-ranking the top-N recommendations (Chapter 5) in order to diversify the results: this approach ensures both properties without interfering with one another. However, there are likely to be more qualities that users seek from their recommendations, and it is likely that they cannot all be optimised independently of one

another.

7.2.2 Beyond Temporal Collaborative Filtering

The themes that emerge from this thesis call for a focus on three key areas in future recommender system research: the *users*, the system *context*, and the *social* aspect of recommender systems.

The methodology that we have used throughout this thesis reflects a *process* that occurs when recommender systems are deployed: users first *rate* content; the ratings are used to compute *predictions*; predictions are ranked to form *recommendations*, and recommendations incentivise users to rate more content. In general, CF research has tended to focus on only two of these three steps. The majority of research (along with the Netflix prize) is dedicated to the problem of computing predictions using ratings. More recently, the importance of ranking (and thus the second step—converting predictions into a ranked list) has emerged, and recommender systems have been evaluated using a variety of information retrieval metrics [ALP+09, Kor08]. However, the last step remains unexplored: given a set of recommendations, why do people rate as they do? Do they rate their recommendations or seek out different content? What affects the way they rate? When confronting the problem of temporal diversity (Chapter 5), we began to see how ratings are not simply reflections of what each user thinks of the *movies* presented to them; the ratings also reflect the users' response to the recommender system and their impression of how well the recommendations are tailored to their needs. A major gap in recommender system research is the focus on the end users' behaviour. While CF has been, for the most part, interesting from the machine learning perspective, it is ultimately an algorithm that has to provide recommendations to people, and further understanding of how the people behave will feed back and improve the algorithms themselves.

A related problem that persists in CF research is that of evaluation: how can researchers demonstrate that their systems are producing "good" results? To date, we have done so by making assumptions of what "good" means: accurate predictions and high precision and recall. In this thesis, we extend that to include, for example, temporally diverse recommendations. We motivated this addition by asking users what they thought of a system that was not temporally diverse. However, what else do users want from their recommendations? The evaluation criteria themselves may be subject to the context in which the recommender system is operating: temporal diversity may make complete sense for a web-based movie recommender system, but may be inappropriate for a system that recommends travel routes to a commuter, unless the diversity is motivated with further reasons (e.g., the current route is congested). On a broader level, systems can be better evaluated if we understand where they will be operating. If future recommender system research focuses on *context*, novel evaluation methods will emerge: for example, to what extent does a travel recommender system on users' mobiles affect their mobility patterns? In other words, are the computed recommendations turning into useful actions?

A final consideration we include is the *social* aspect. There is an overlap between social networks and collaborative filtering; in fact, we were able to draw from social network analysis techniques in order to examine how similarity graphs are structured (Chapter 3). In doing so, we claimed that CF rating data represents an *implicit* social network between the system users, because what one person rates affects others' recommendations. The implication here is that CF ratings are related to one another;

in fact, there may be a causal relationship between users' ratings. These relationships are difficult to understand since the links between users remain hidden; however, future research based on combined (social network/content ratings) datasets will be able to investigate this link further and use it to improve the recommendations each user is given. Recent web-based companies are already gathering data that will serve this purpose: for example, Rummble[1] gathers users' *social network* and *ratings* on different locations around the world.

[1] http://www.rummble.com