Research Lines in Secure and Robust Recommender Systems

Neil J. Hurley and **Guénolé C.M. Silvestre**
School of Computer Science and Informatics
University College Dublin
Belfield
Dublin 4, Ireland

Abstract

Monitoring security and trust in on-line personalised recommendation systems is now recognised as a key challenge. Noisy data, or maliciously biased data, can significantly skew the system's output. This paper outlines our research goals, which aim to tackle this issue along a number of lines. Game theoretic techniques are applied to determining bounds on the effect of robustness attacks on recommender systems. Graph theoretic techniques are used to analyse the dataset structure and identify influential users in the application user-group, for filtering purposes.

A user profile database is a key component of most on–line e–commerce systems. The database stores user information such as the history of access patterns to the site, product preferences and profile information such as age, profession and so on. This information is used to drive personalisation software to provide individually customised interfaces for each customer that, for example, direct customers to products they are likely to have an interest in. In many cases these databases are open in the sense that they are continuously updated by information gathered from customers. From the site manager's point–of–view this is good, since it allows for system responses to be dynamically improved as more information about the customer is obtained. However it is also a security concern, since it is often impossible or intractable to confirm the accuracy of customer–inputted information and an accumulation of inaccurate information can seriously degrade the service. From the customer's perspective, issues of privacy are of significant concern and generally customers are uncomfortable with the collection of detailed personal information. Peer–to–peer systems, such as the eBay auctioning service, involve blind interactions between individuals over the Internet, including agreements to trade. For such systems, the issue of trust between the parties engaged in the transaction is of major concern and much effort has been made to construct reputation reporting algorithms that can automatically calculate the trustworthiness of individuals. Such algorithms typically depend on the user community self–monitoring through a process of individuals rating their satisfaction with the other parties that they have traded with. Hence, such algorithms depend on an open database of information gathered dynamically through user interaction.

Originally, personalisation algorithms were evaluated in terms of the quality of service they provided, assuming that databases were accurate. Database quality was measured only in terms of the sparseness of the available dataset, with the assumption that system performance should improve as more data becomes available. However, we have carried out work (O'Mahony, Hurley, & Silvestre 2004b; 2004a; 2005) that has demonstrated that not only is the data quality an important concern to personalisation algorithms, relatively small quantities of maliciously tailored data, entered into systems through open user interfaces, can drastically degrade the performance of the system.

This work, along with the work of others (Lam, Frankowski, & Riedl 2005; Burke, Mobasher, & Bhaumik 2005), is part of a general upsurge in interest in the community into issues of trust and security in personalisation software.

In the last number of years, we have carried out an intensive investigation of robustness in the context of the collaborative filtering recommendation algorithm. We have shown the degree to which robustness is a problem and have proposed some ways to counter the robustness problem. All solutions rely on filtering *unusual* data from the database. However, this is in opposition to the general philosophy of personalisation. Customers' individuality is determined by how they differ from the norm. There is therefore a danger that, particularly as databases scale to large numbers of customers, filtering algorithms employed to increase the security and robustness of the system may in fact reduce the system's ability to provide a truly personalised service. Hence, the general question that this research will attempt to answer is how to tackle robustness and security *while maintaining* personalised performance.

In summary, our research addresses the problem of developing personalisation algorithms that are:

- **robust**, in the sense that they can operate effectively with noisy data;
- **secure**, in the sense that they cannot be manipulated by potentially malicious end–users;
- **user–centric**, in that they take into account privacy re-

quirements of the customer base;

- **scalable**, in that they can deal efficiently with large datasets; and
- **trust–aware**, in that they take into account the trustworthiness of the data available to them.

How is the question being Addressed?

Recommender systems (RS) are subject to a suite of malicious attacks, such as denial-of-service attacks, that apply to any client-server system. However, we focus not on such system-level attacks, but rather on attacks that compromise the quality of the personalisation data exploited by the system in order to make user-centred recommendations. We have shown (O'Mahony, Hurley, & Silvestre 2005) that a small number of false identities can be used to significantly bias the RS output (now generally referred to as *shilling* attacks). Thus a book author can, through the creation of a set of false identities, all promoting his work, bias the RS to output, on average, higher ratings for his books (a so-called, product promotion or *push* attack). Further, we have demonstrated that profile filtering (or *obfuscation*) strategies to detect or remove the influence of such false identities, lead to an arms race, in which, an attacker aware of the obfuscation strategy, can modify the attack to circumvent it (O'Mahony, Hurley, & Silvestre 2004a). Much remains to be done to understand the limits of such attacks and their implications for RS systems. We are focusing our research along the following lines:

Cost Benefit Analysis One of the out-standing research questions concerning *shilling* attacks, is whether they are powerful enough to be carried out in a cost-effective manner. In other words, although effective attacks that are small with respect to the database size can be constructed, it is not clear whether an attacker can accrue real profits through carrying out such attacks. Adopting a simple model of the recommendation process, we have recently (Hurley, Mahony, & Silvestre 2007) given initial results in this regard, but further work is required. Our approach is to compute (pre- and post-attack) the probability, $\rho(u)$ that a user u *eventually* purchases an attacked item. Writing this probability in terms of the probability q that the item appears in the recommended set R_j at transaction j and making some assumptions of independence, we get

$$\rho(u) = \sum_{n=0}^{\infty} \int_0^1 \left(1 - (1 - pq)^n\right) \phi(n) f_q(q) dq \,, \qquad (1)$$

where $f_q(q)$ is the probability distribution function of q, $\phi(n)$ represents the probability that a user carries out n transactions in a given time period and p represents a probability of acting on the recommendation. Modelling $\phi(n)$ as a poisson process, this model leads to near linear growth in expected income from an attack, with growing database size (in terms of numbers of users). While it represents good news to an attacker, it depends on some difficult-to-measure parameters and some unrealistic assumptions. A more exact analysis is possible. The approach that we are following is to adopt a probabilisitc model, in which a given user-profile dataset is understood as a realisation of the stochastic process, $D(t)$, the $U \times I$-dimensional user-profile matrix at time t, where U is the number of system users and I is the number of items. Imposing input probability distributions on user-profiles pre- and post-attack, we are seeking to derive the distribution of the output of the prediction algorithm, pre- and post-attack and hence, as above, derive a probability that an item appears in the recommender set. Given time-stamped ratings data (available in datasets such as Movielens and NetFlix), it is finally possible to compare statistically the actual user actions at time t, with the recommendations that the system can offer at time t and thus derive a model of how users act on recommendation output.

Nevertheless, this model only examines the financial benefit that can accrue to an attacker. In reality, it is also important to consider the *risk* associated with an attack, if being detected has severe consequences. Considering that the recommendation algorithm designer wants to assure the best performance of the recommender system and an attacker seeks to surreptitiously destroy the recommendation quality, it is plausible that this scenario can be tackled in a **game theoretic** framework. The shilling attacks proposed by our previous work can be distinguished from other proposed attacks by the fact that we assume that the attacker has full knowledge of the underlying recommender algorithm. Although this may seem unrealistic in practice[1], the approach follows Kerchoff's Principle from cryptography that states that security should not depend on the secrecy of the underlying algorithm. It allows us to find lower bounds on performance in a system under attack.

If an appropriate game model yields an equilibrium, then it will give the performance of the system at a point where there is no advantage for the attacker or algorithm designer to modify their strategies. Even if an equilibrium does not exist, the minimax and maximin problems defined in the game will yield upper and lower bounds to the system performance under attack. The formulation of an appropriate game is a matter of research. The following provides a simple example to clarify the approach. The algorithm designer makes predictions using some set of weights, which we denote as $\boldsymbol{\beta}$. These weights encapsulate the different algorithm variations available to the designer and may reflect, for example, the designer's trust in particular user profiles, as well as similarities to the active user for whom a rating is predicted. The attack may be simply modelled as (possibly biased) noise insertion into the database ratings, where the attack strength is measured in terms of σ_n the root mean-squared rating error inserted by the attacker. To define a game, it is necessary to identify the *payoff* function. In this scenario, the payoff function is recommendation accuracy, as measured by the mean squared prediction error, $e(\boldsymbol{\beta}, \sigma_n)$ which the algorithm designer wants to minimise and the attacker wants to maximise. This corresponds to a game in which the attacker is simply attempting to destroy overall system performance rather than push a particular item. Thus the game consists of successive maximisation

[1] We have shown that attacks are still effective when the available knowledge is reduced.

and minimisation of e by the attacker and algorithm designer i.e. $\min_{\boldsymbol{\beta}} \max_{\sigma_n} e(\boldsymbol{\beta}, \sigma_n)$. The game has a pure equilibrium if the minimax solution equals the maximin solution for some optimal $\boldsymbol{\beta}^*$ and σ_n^*.

Assume that a prediction $p_{a,k}$ is made for a user a and item k as a weighted sum, with user-user similarity weights $\beta_{a,u}$, of the ratings given by the other users. Assume, for a training set of U users and I items, all ratings $r_{u,i}$ are available and are used by the algorithm designer to select user-user similarity weights $\beta_{u,v}$, by minimising the prediction error in a 'leave-one-out' analysis (i.e. predictions $p_{u,i}$ are made for each user u and item i in the training set, using all the available data, except the rating $r_{u,i}$.) Hence we have the matrices $P = \{p_{a,k}\}$, $R = \{r_{a,k}\}$ and $B = \{\beta_{a,k}\}$ such that $P = BR$ and $\mathrm{diag}(B) = 0$. The mean-squared prediction error can then be written in terms of the frobenius matrix norm as

$$\begin{aligned} \sum_{a,k} (p_{a,k} - r_{a,k})^2 &= \|P - R\|^2 \qquad (2) \\ &= \mathrm{Tr}((BR - R)^T (BR - R)) \end{aligned}$$

The algorithm designer's object is to minimise this expression wrt the matrix weights B. For simplicity, relaxing the constraint on the diagonal to $\mathrm{Tr}(B) = 0$, the constrained optimisation problem can be solved by differentiating wrt B to get

$$B^* = I - \alpha (RR^T)^{-1} \qquad (3)$$

such that

$$\alpha = \frac{U}{\mathrm{Tr}((RR^T)^{-1})} \qquad (4)$$

and hence

$$e^* = \frac{U^2}{\mathrm{Tr}((R^T R)^{-1})}\,. \qquad (5)$$

An attack modifies the ratings matrix to $R' = R + \Upsilon$, where Υ represents the noise that is added to each rating. Given the original ratings matrix, and examining (5), an attacker can maximise the error by minimising the trace term, which leads to $\Upsilon = -\gamma R$ for some γ, which is bounded using the attack strength σ_n.

Of course, in practice, a real attacker does not have access to the original ratings and can at best estimate them. Moreover, an attacker will not be able to add arbitrary noise to all ratings in the database, and indeed operational recommendation algorithms do not perform a simple weighted sum as described above. Nevertheless, the above scenario gives some indication of how the game theoretic approach can allow the strategies available to the algorithm designer and the attacker to be clearly formulated and analysed. In our research we are investigating game-theoretic models that are closer to reality with the goal of computing bounds on pay-offs that can be interpreted from a cost-benefit point-of-view.

Graph Theoretic Analysis An analogy can be drawn between security of RS systems and security of computer networks in general. The topology of the Internet has been recognised to be scale-free, leading, from a security perspective, to a interconnection network that is often referred to as *robust but fragile* (Li *et al.* 2005). It is robust in the sense that a random removal of a node from the network will not compromise its performance, but fragile, because a targeted attack that removes so-called *hubs* can quickly bring down the entire system. A hub is a node of high out-degree, thus connecting to many other nodes in the network and creating small-length paths between pairs of nodes. Shilling attacks on an RS can be interpreted in terms of an insertion of hubs into the user-profile dataset. Attack profiles are designed to have high similarity with as many users as possible. Hence in the graph of user-user similarities, nodes representing attack profiles have high degree (see Fig. 1). While this observation provides a simple way to detect potential attack profiles (or at least highly influential profiles), it also suggests that a structural analysis of the recommendation dataset can lead to a deeper understanding of the performance of the system. Thus following research lines such as (Li *et al.* 2005) in the context of computer networks, we are addressing the question of robustness of RS in terms of the underlying structure of the dataset. Early investigations along this line have uncovered a number of interesting research questions. For example, although much RS research has focused on the *sparseness* of the dataset, none has investigated deeper features of the structure. In Fig. 2, a number of different RS datasets are shown, all with the same spareness, but different interconnection structure. It can be concluded that structural features other than sparseness can have a greater impact on performance. Using such graph theoretic techniques, as well as addressing the robustness question, it may be possible to produce *quantitative* measures of the *extent* to which a particular RS algorithm can work for a particular user domain.

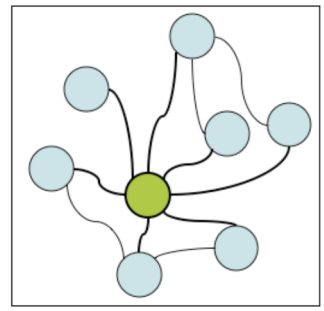

Figure 1: The central node represents an attack profile in the user similarity space of a collaborative recommendation dataset. By carefully choosing the make-up of the profile, it is possible to ensure that the profile has a high similarity to many genuine profiles in the dataset. Graphically, this corresponds to a high-degree node, or *hub*.

Furthermore, it may be possible to develop *synthetic* datasets that properly model user behaviour – providing a useful tool for RS researchers, particularly given the privacy issues surrounding the use of real user data. This amounts to closing the gap between the random 'srand' dataset and 'Movielens' in Fig.2. Although the gap appears small, early analysis has shown that closing this gap is a non-trivial task.

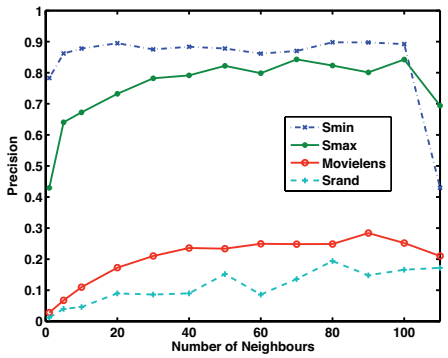

Figure 2: Precision measure of the performance of the collaborative filtering recommendation algorithm, applied to a number of datasets. Each dataset has exactly the same sparsity, but a different structural arrangement.

Attacks succeed because attackers, with a small amount of domain knowledge, can tailor highly influential profiles. Much recent interest has focused on the analysis of social network graphs, from the point-of-view of identifying user communities and determining influential individuals within those communities. Such techniques can be used to identify attackers or indeed to simply identify genuine but overly influential profiles, so that filtering algorithms, targeting such profiles can help to improve overall system performance in terms of accuracy and robustness. Strategies for community identification are based on spectral graph partitioning schemes (White & Smyth 2005), that optimise a utility function measuring the connectedness of potential communities. Alternatively, graph theoretic measures such as *betweenness* – defined for each node k as the number of pairs of nodes with the shortest path among them passing through k normalized by the total number of pairs of vertices – provide ways of identifying influential members in the community graph.

We are focusing on applying these measures and techniques to identifying influence in the recommendation dataset. In this regard, we see a number of challenges and possibilities for extending the current state-of-the-art in community analysis. One challenge is that the graphs typically analysed by these social network algorithms are unidirected user-user graphs. In the recommendation problem, it is more tractable to deal directly with the *bipartitie* user-item graph. Another avenue to extend the state-of-the-art is to investigate strategies that allow for *soft* membership of communities (allowing individuals to belong to more than a single community).

Conclusion

In this research note, we have outlined some research directions towards the development of secure and robust recommendation systems, focussing particularly on how game theoretic analysis and graph theoretic analysis can provide a deeper insight into attack strategies that have been developed along heuristic lines in recent years.

Acknowledgements

This work is supported by Science Foundation Ireland.

References

Burke, R.; Mobasher, B.; and Bhaumik, R. 2005. Limited knowledge shilling attacks in collaborative filtering. In *Proceedings of the Workshop on Intelligent Techniques for Web Personalization (ITWP'05)*.

Hurley, N.; Mahony, M. O.; and Silvestre, G. 2007. Attacking recommender systems: The cost of promotion. *IEEE Intelligent Systems, Special Issue on Recommender Systems* 22(3).

Lam, S. K. T.; Frankowski, D.; and Riedl, J. 2005. Do you trust your recommendations? an exploration of security and privacy issues in recommender systems. Technical report, GroupLens Research, Computer Science and Engineering,University of Minnesota.

Li, L.; Alderson, D.; Tanaka, R.; Doyle, J. C.; and Willinger, W. 2005. Towards a theory of scale-free graphs: Definition, properties,and implications (extendedversion). Technical Report CIT-CDS-04-006, Engineering & Applied Sciences Division, California Institute of Technology, Pasadena, CA, USA.

O'Mahony, M.; Hurley, N.; and Silvestre, G. 2004a. An evaluation of neighbourhood formation on the performance of collaborative filtering. *Artificial Intelligence Review* 21(1):215–228.

O'Mahony, M.; Hurley, N.; and Silvestre, G. 2004b. Utility-based neighbourhood formation for efficient and robust collaborative filtering. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC'04)*.

O'Mahony, M.; Hurley, N.; and Silvestre, G. 2005. Recommender systems: Attack types and strategies. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI'05)*, 334–339. AAAI Press, Pittsburgh, Pennsylvania, USA.

White, S., and Smyth, P. 2005. A spectral clustering approach to finding communities in graphs. In *SIAN International Conference on Data Mining*.