Protecting the Privacy of Observable Behavior in Distributed Recommender Systems

Douglas W. Oard*, Anton Leuski† and Stuart Stubblebine‡

1 Introduction

Implicit and explicit evidence can be thought of in a unified framework as "observable behavior;" explicit rating is merely one type of behavior [1]. Adequate protection of privacy is a prerequisite to the use of observable behavior (a necessary, but not a sufficient condition—a willingness to help shape the information space is also required, for example). Privacy can be protected in cases where the behavior is intended to be public (e.g., building Web links) or when users believe that adequate safeguards exist (e.g., Amazon's use of purchase behavior to make recommendations). We are interested in exploring distributed techniques that can be used in cases where centralized services are unable to establish acceptable privacy safeguards.

In the remainder of this paper we use U = *user*, I = *item*, B = *behavior*, R = *recommendation*, and F = *feature*. Centralized recommender systems based on implicit feedback often map from a $U \times I \times B$ array of observations to a $U \times I$ matrix of recommendations. This matrix is then used to form either an $I \times I$ matrix of item similarity that can be used as a basis for "cross-selling," or a $U \times U$ matrix of user similarity that can be used to find users with similar tastes, with which recommendations can then be constructed using the $U \times I$ matrix.

In distributed systems, sharing information about users is potentially problematic, even when pseudonyms are used, because side information could serve to pinpoint the identity of an individual (e.g, there may be only one pilot who lives in College Park that does information retrieval research) [2]. Some distributed ways of building the $I \times I$ matrix may not suffer from this weakness, but that matrix models preferences as scalar relationships, while true preference relationships are an individualized aggregate of multiple factors. We therefore define a more abstract $I \times F$ matrix to be shared and used as a basis for making recommendations.

At this point, we make no commitment to the meaning of individual features. We require only that users be able to compute a personally useful $I \times R$ matrix from the $I \times F$ matrix, and that they be able to update the $I \times F$ matrix based on their personal (and private) $I \times B$ matrix. A restricted case of this would be to treat items as independent and compute mappings from B to F and from F to R. For example, B = (read 15 seconds, forwarded to boss, saved) might map to F = (related to information retrieval, high quality), which in turn might map to R = (high interest at work, low interest at home). Each community must agree on a common space for F, but the nature of B and R need not be completely standardized.

*College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, oard@glue.umd.edu

†University of Southern California Information Sciences Institute, 4676 Admiralty Way, Suite 1001, Marina Del Rey, CA 90292, leuski@isi.edu

‡Stubblebine Research Labs, LLC, 8 Wayne Blvd., Madison, NJ 07940-1529, stuart@stubblebine.com

In this extended abstract we introduce the framework described in the context of security issues for protecting the privacy of individuals that contribute to the incremental creation of a shared public $I \times F$ matrix. We conclude with a brief description of what we see as the next steps for exploring this question. Our intent is to identify some key research challenges that must be addressed before such systems can be deployed in a much broader range of applications than is presently the case.

2 Attack Scenarios

In this section we describe three ways in which the integrity of a distributed system based on observable behavior might be compromised, and we identify some approaches to mitigating these risks. The first two scenarios deal with controling what inferences the adversary might learn from available information.

2.1 Scenario 1: Low Entropy Attack

The basic idea is that an adversary uses the $I \times F$ matrix along with some side information to reduce his uncertainty about the $I \times B$ matrix for some user to below an acceptable threshold [2].

For example, consider the case in which each feature is uniquely associated with a single user):

- F_1: Liked by Doug
- F_2: Liked by Anton
- F_3: Liked by Stuart

In this case, if we have as side information the interpretation of F_1, F_2, and F_3, we can create a $U \times I$ matrix (where each element of that matrix is B, an explicit rating) by simply replacing F_1, F_2 and F_3 with Doug, Anton and Stuart respectively. We can thus associate a behavior with a specific user.

Formally, we seek to ensure that the cross-entropy between the $I \times F$ matrix and any element in the $U \times I \times B$ array can be made sufficiently large under some specified set of assumptions about the side information that is available to the adversary.

Possible approaches to protecting privacy against such an attack might include:

1. Precluding the ability to benefit from contributions to the $I \times F$ matrix until some precondition is met—such as a sufficiently large number of contributions have accumulated. However, this approach extends the duration of the cold start problem.

2. Limiting the privacy protection afforded through technical means to a subset of users that "hide behind" the activities of the rest of the community. This might be possible if the privacy of other users be adequately assured by other means (e.g., legal protection against misuse of information), or if a sufficient number and sufficient diversity of users that lack privacy concerns are available.

2.2 Scenario 2: Matrix Difference Attack

The basic idea is that an adversary sees the $I \times F$ matrix before and after some user updates it, and the adversary uses that to create an $I \times \Delta F$ matrix that can be used to reduce the adversary's uncertainty about the user's $I \times B$ matrix to below an acceptable threshold.

For example, consider the case in which the features have the following interpretations:

- F_1: Information is of interest to people planning to file a patent
- F_2: Information is of interest to people planning a surprise party
- F_3: Information is of interest to people that test positive for HIV

There are many ways that the $I \times F$ matrix might be structured, but consider the case where it merely contains counts of the number of users that exhibited behavior that would lead you to infer a positive value for the feature. Then the $I \times \Delta F$ matrix can be computed as $(I \times F)_{after} - (I \times F)_{before}$. If we assume that the adversary knows which user just updated the $I \times F$ matrix, then the adversary can determine information that reduces the uncertainty about the user's behavior (e.g., which items they rated highly, if the ratings were explicit). Note that in this case it is also possible to infer information that extends beyond the user's behavior (because of the side information about the meaning of the features) that could not be deduced without access to the evidence about their behavior.

Formally, we seek to ensure that the cross-entropy between (a) two instances of the $I \times F$ matrix that are related by a known sequence and available to the same observer and (b) any element in the $U \times I \times B$ array can be made sufficiently large under some specified set of assumptions about the side information that is available to the adversary.

Possible approaches to protecting privacy against such an attack might include:

1. Restricting the ability of an observer to gain access to the contents of two matrices that would yield an exploitable matrix difference. For example, we could enforce a routing scheme in which the source or the destination is randomly selected from a sufficiently large subset of users. The benefit here is that the adversary doesn't know who to target or collude with in order to compromise the security goal. Ideally, the adversary should not be able to correlate which users have or have not contributed to the matrix based on observable system information [4].

2. Encrypting the $I \times F$ matrix in such a way that an adversary that has any two such matrices would gaining no more information than they could gain from the more informative of the two. Such an approach would require assuming limited computational effort by the adversary, since with enough effort the adversary could explore the full space of user characteristics that form the basis for the transformation from $I \times F$ to $I \times R$, thus ultimately discerning additional information from the second $I \times F$ matrix.

2.3 Scenario 3: Identity Integrity Attack

The basic idea here is that an adversary registers myself as more than one user to spam the $I \times F$ matrix. While not actually a risk to privacy, the risk of an identity integrity attack can be exacerbated by any degree of anonymity that we might impose to accommodate Scenario 1. Moreover, if adversaries can potentially register an unlimited set of pseudonyms, that proliferation could compromise some countermeasures that might be considered for Scenario 2. For example, if we use pseudonyms to protect privacy, consider the case in which an adversary registers for enough pseudonyms to dominate the information flow into the system, this gaining the ability to reconstruct updates to the $I \times F$ matrix by individual users.

A possible approach to protecting identity integrity would be use of a registration service to issue cryptographic credentials (e.g., a digital certificate that can serve as a pseudonym) after leveraging some existing authentication infrastructure. A key property of such a service is that a single user should be able to obtain at most one active credential [3]. An existing identity authentication

service (e.g., based on credit reports) can serve as the basis for that assurance if the integrity of the base service is adequately protected through other means. Attacks on the registration service itself can be mitigated using a secure quorum scheme that is robust against introduction of a bogus identity server.

3 Future Work

In this extended abstract we introduced a framework for sharing the information about observable behavior and then proposed a set of properties that are necessary (and, we hope, sufficient) to protect the privacy of individuals that contribute to the incremental creation of a shared public $I \times F$ matrix.

This is a work in progress, and at present we have more questions than answers. The next step will be to craft a protocol for information exchange that we can test for the vulnerabilities that we have identified and around which we can build the needed security infrastructure. To do that, we will need to define the contents of the $I \times B$ matrix, the way in which the $I \times F$ is updated using the $I \times B$ matrix, the way the $I \times R$ matrix is computed from the $I \times F$ matrix, and the way the $I \times F$ matrix is exchanged around among the participants. To do so, we will need to focus on a specific application where B, F and R take on concrete meanings. Only after trying this for several applications do we expect that we would be able to discern general principals that could be offered as prescriptions for the protection of privacy in such systems. This is more than one small team could hope to accomplish on its own, so we are interested in discussing this framework at the workshop to see if there is merit in our general approach and if there is interest in a concerted attack on this problem.

References

[1] D. W. Oard and J. Kim. Modeling information content using observable behavior. In *Proceedings of the 64th Annual Conference of the American Society for Information Science and Technology*, pages 481–488, Washington, 2001.

[2] N. Ramakrishnan, B. J. Keller, B. J. Mirza, A. Y. Grama, and G. Karypis. Privacy risks in recommender systems. *IEEE Internet Computing*, pages 54–62, nov 2001.

[3] S. G. Stubblebine and P. F. Syverson. Authentic attributes with fine-grained anonymity protection. In *Financial Cryptography*, pages 276–286. Springer-Verlag, 2001.

[4] S. G. Stubblebine, P. F. Syverson, and D. M. Goldschlag. Unlinkable serial transactions: protocols and applications. *ACM Transactions on Information and System Security*, 2(4):354–389, 1999.