

# Secure Personalization

## Building Trustworthy Recommender Systems

Robin Burke & Bamshad Mobasher, DePaul University, Chicago, Illinois



Bamshad Mobasher



Robin Burke

## Attacks and Remedies in Collaborative Recommendation

The goal of this project is to explore the vulnerabilities of recommendation and personalization systems in the face of malicious attacks, explore techniques for enhancing their robustness, and examine methods by which attacks can be recognized and possibly defeated.

### Collaborative Recommendation: Highly Vulnerable

Collaborative filtering (CF) recommendation is commonly used in e-commerce. Users' preferences are represented by profiles that consist of ratings for products. Such systems identify peer users with similar tastes and extrapolates their ratings as likely to be similar to the target user. The standard formulation of CF is highly vulnerable to attack. An attacker can insert a large number of fake profiles to bias the system's recommendations for or against certain products.

### Securing Collaborative Recommendation

Our research has focused on the development of more robust algorithms as well as methods for attack detection.

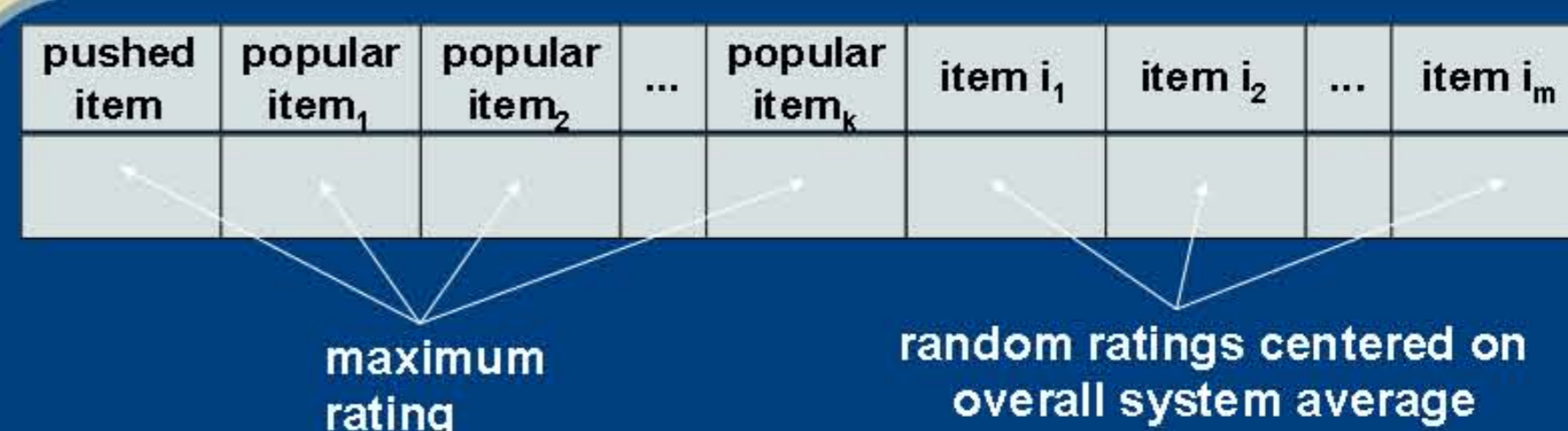
**Algorithms:** We have explored the spectrum of possible attacks against recommendation systems, and developed models characterizing these attacks and their impacts. We have examined a range of recommendation algorithms including user-based, item-based and model-based collaborative recommenders, and developed a **more robust hybrid algorithm**, combining collaborative with content-based and knowledge-based techniques.

**Detection:** One approach to attack detection is based on classification learning:

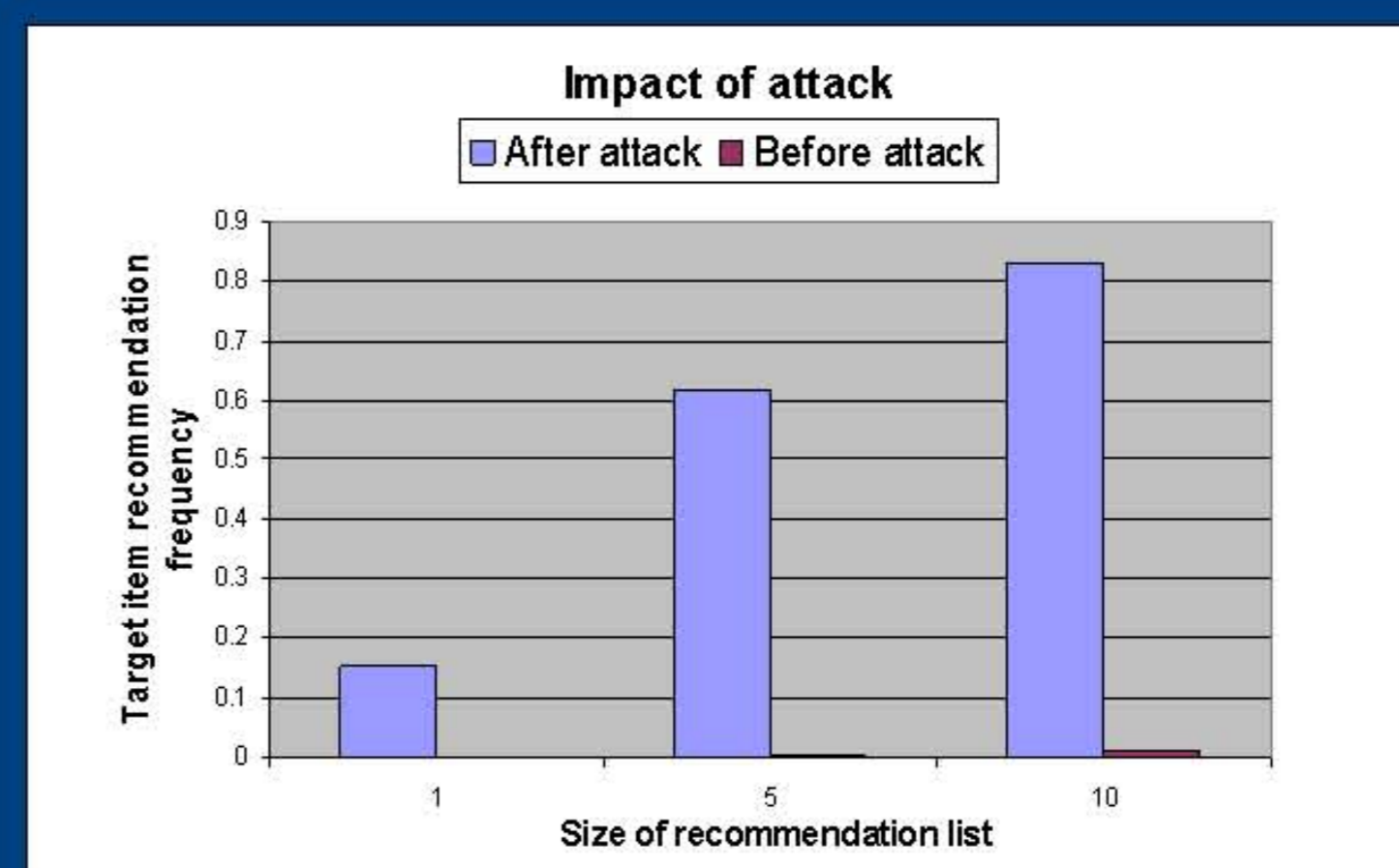
- Create attributes that capture characteristics of suspicious profiles
- Use attributes to build classification models
- Apply model to user profiles to identify and discount potential attacks

#### Two Types of Detection Attributes

- **Generic** – Capture statistical characteristics of typical attack profiles, e.g., differences of profile's ratings from mean rating on each item, or average correlation of the profile's  $k$  nearest neighbors.
- **Model-specific** – Based on characteristics of specific attack models, e.g., the rating variance between the highly rated selected items versus the items with low ratings.

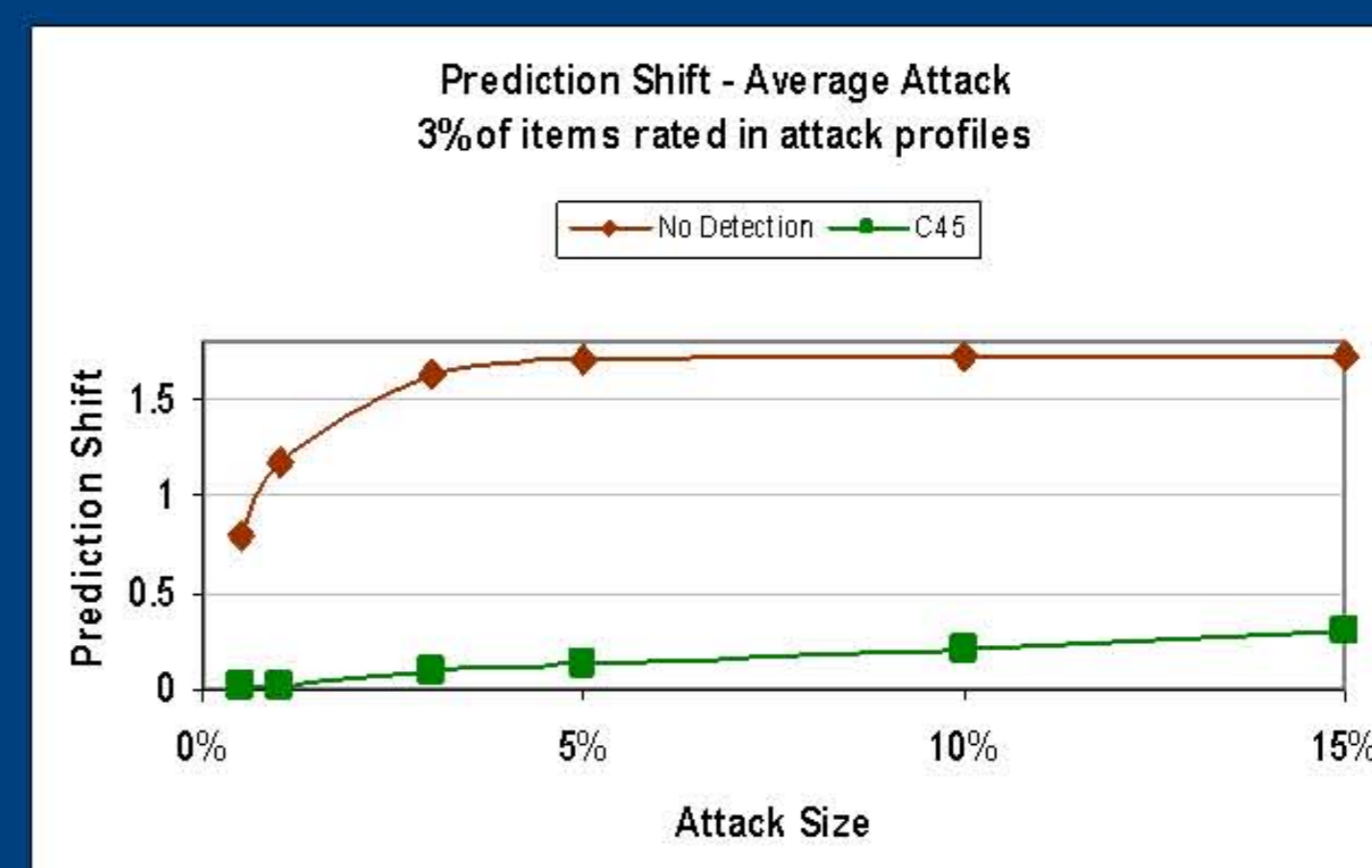


Example: **Bandwagon attack**. Profiles use popular items to **push** a target item.



After the attack, item appears in 80% of top 10 recommendation lists.

Example of attack detection results: Using C4.5 Decision Tree model based on generic and model-specific attributes.



**Prediction shift** (rating scale of 1-5) due to attack, with and without detection.

## Summary of Findings To Date

**Theoretical Framework for Attack Model:** Has enabled the systematic study of profile injection attacks and their properties.

**New Algorithms:** Significant advantage for hybrid recommendation (knowledge-based & collaborative). Also, certain model-based algorithms are more robust with minimal loss in prediction accuracy.

**Attack Detection:** Profile classification approach based on attack model specific attributes has led to highly effective approach for detecting most attack models.

**Anomaly Detection:** Classify items (as being under attack). Not dependent on known attack models, and can determine which type of items are most vulnerable to which types of attacks.

## Real world impact

Examples of "shilling" are well-known in the industry. When a security breach caused the release of reviewers real names, Amazon.com discovered that many authors were pseudonymously reviewing their own books. One e-commerce site found that a manufacturer was using off-shore sub-contractors to enter hundreds of favorable reviews of its products. Secure recommendation algorithms would increase user trust and allow site operators to realize greater benefits from their recommender systems.

<http://maya.cs.depaul.edu/~mobasher/cgi-bin/view-pubs.pl?CID=SP>