Towards a Comprehensive Testbed to Evaluate the Robustness of Reputation Systems against Unfair Rating Attacks

Athirai Aravazhi Irissappane, Phd student, IMI & SCE, NTU
Supervisor: Assistant Professor Jie Zhang
Co-Supervisor: Professor Nadia Magnenat Thalmann

January 15, 2013

Overview

1 Introduction and Motivation

Overview

1 Introduction and Motivation

2 Proposed Testbed

Overview

1. Introduction and Motivation
2. Proposed Testbed
3. Case Studies

Overview

1. Introduction and Motivation
2. Proposed Testbed
3. Case Studies
4. Conclusion

Introduction

State of Most Users in an E-Transaction

- Which buyers/sellers should I trust?
- Which seller will give me the best quality product?
- To which other buyers should I ask advice about the sellers?

....

Reputation Systems

1. Facilitate trust in internet interactions
2. Calculate the reputation of an agent using the ratings given by the other entities to the agent

However, reputation systems are affected by **unfair rating attacks** from dishonest entities!

How to Combat Unfair Ratings?

Several approaches have been proposed by researchers which,

- Detect unfair rating attacks
- Accurately evaluate the reputation of sellers

Some examples of such reputation systems with unfair rating detection approaches are:

- BRS [Whitby and Jøsang et al., 2004]
- TRAVOS [Teacy et al., 2006]
- Personalized [Zhang and Cohen, 2008]
- WMA [Yu and Singh, 2003]
- iCLUB [Liu et al., 2011]

Motivation

How Reliable the Existing Reputation Systems are?

- They are evaluated using methods of the **authors' own devising**
- The evaluation is mainly **simulation based** which may not reflect the real environment
- Lack of extensive performance evaluation in **complex attacking scenarios** like the collusion attacks

Testbeds to evaluate the performance of the reputation systems exist, but have certain limitations.

Existing Testbeds

- **ART** [Fullam et al., 2005]: The ART testbed specification is an artwork appraisal domain where appraisers need to buy artwork about which they may have limited knowledge.

  Limitations:

  a. Simulation based
  b. No specific robustness evaluation metrics
  c. Integration with the testbed is quite challenging

- **TREET** [Kerr and Cohen, 2010]: TREET is a testbed which models a general e-marketplace scenario. It supports both centralized and decentralized reputation systems and allows collusion attacks to be implemented.

  Limitations:

  a. Simulation based
  b. Not specifically designed to evaluate unfair rating detection approaches

Our Objective is to Develop a Comprehensive Experimental Testbed with the Following Characteristics:

- Perform detailed evaluation of the reputation systems to specifically deal with **unfair rating attacks**
- Conduct **robustness** evaluation of the various approaches
- Conduct reliable evaluation based on **real data**

Proposed Testbed

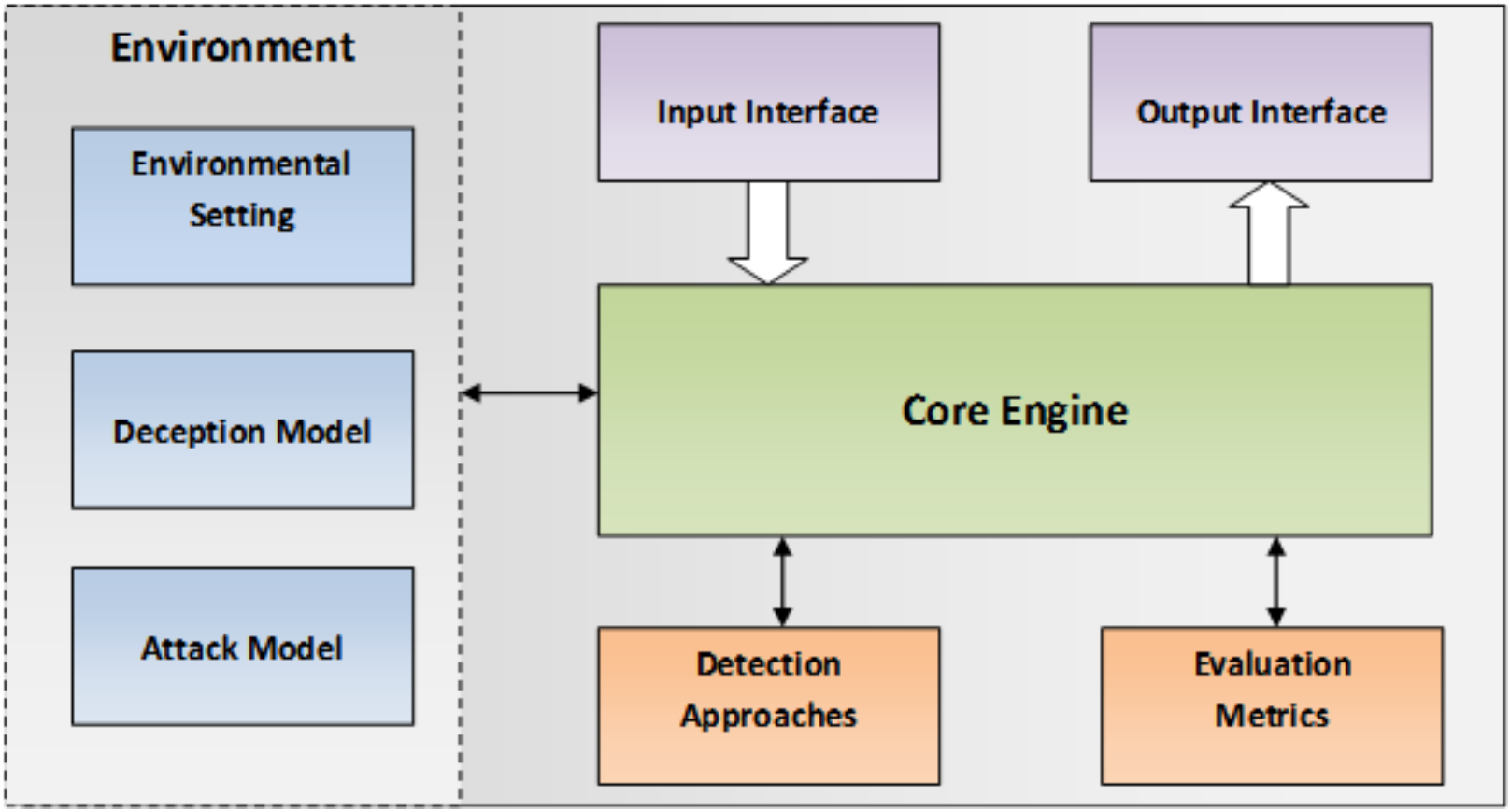

Figure: High Level Architecture of the Testbed

Major Components of the Testbed:

- **3 Kinds of environments**:
  1. **Simulated environment:** Environment is entirely based on simulations
  2. **Real environment with simulated attacks:** Data is collected from real environments. Various attacks are simulated to provide unfair ratings which are added to the real data
  3. **Real environment with detected "ground truth":** Spam review detection tools are used to detect spam reviews in the collected real data

- **Detection approaches**: Represents the reputation systems to be compared and evaluated against each other
- **Input interface**: Helps to configure different environmental settings
- **Output interface**: Provides the necessary visualization tools to easily comprehend the results
- **Core engine**: Responsible for managing all the related interfaces of the system

- **2 Novel robustness metrics**:
  - The robustness $\mathcal{R}(Def, Atk)$ of a reputation system (defense, Def) against an unfair rating attack model (Atk) is :

$$\mathcal{R}(Def, Atk) = \frac{|Tran(S^H)| - |Tran(S^D)|}{|B^H| \times |Days| \times ratio}$$

    If completely robust, honest duopoly sellers will get more number of transactions. The greater the value $\mathcal{R}(Def, Atk)$ is, the more robust Def is against Atk.

  - The number of unfair ratings required by attackers to change a target's reputation, because a more robust reputation system costs more efforts from attackers

Case Study 1: Simulated Environment

The Environment is Entirely Based on Simulation

- **Simulation settings:**
  - No.of honest, dishonest duopoly sellers = 1 each
  - No.of honest, dishonest common sellers = 99 each
  - No.of honest buyers/advisors = 28 (non-Sybil based attack), 12 (Sybil based attack)
  - No.of dishonest buyers/advisors or attackers = 12 (non-Sybil based attack), 28 (Sybil based attack)
  - No.of simulation days = 100
  - Ratio of duopoly sellers' transactions to all = 0.5

- **Detection approaches:** BRS, iCLUB, TRAVOS, WMA and Personalized approach
- **Attack models**: Constant Attack, Camouflage Attack, Whitewashing Attack, Sybil Attack, Sybil Camouflage Attack and Sybil Whitewashing Attack
- **Evaluation metric:** The reputation systems are evaluated based on the first robustness metric (difference in transaction volume between honest and dishonest duopoly sellers)

Case Study 1: Experimental Results

Table: **Robustness of Reputation Systems Against Attacks**

	Constant	Camouflage	Whitewashing	Sybil	Sybil Cam	Sybil WW
BRS	0.87±0.03	0.89±0.02	-0.18±0.07	-0.99±0.08	-0.47±0.07	-0.30±0.07
iCLUB	0.98±0.03	0.99±0.03	0.79±0.14	0.21±0.32	**0.94±0.10**	**0.20±0.29**
TRAVOS	0.97±0.02	0.82±0.03	0.87±0.03	0.16±0.09	-0.57±0.07	-0.98±0.07
WMA	0.89±0.04	0.69±0.04	-0.95±0.08	**0.82±0.06**	0.63±0.08	-0.98±0.07
Personalized	**0.99±0.03**	**0.99±0.03**	**0.98±0.03**	0.74±0.45	**0.94±0.08**	-1.00±0.08

*Sybil Cam: Sybil Camouflage Attack; Sybil WW: Sybil Whitewashing Attack

The table shows the $mean \pm std$ robustness values of the reputation systems over 50 simulation runs

- None of the reputation systems are completely robust against all the attacks
- iCLUB obtains 2 best results for Sybil Camouflage and Sybil Whitewashing attacks
- WMA obtains 1 best result for Sybil
- Personalized obtains 4 best results
- All the reputation systems are robust against Constant attack
- None of reputation systems are completely robust against Sybil Whitewashing attack

Case Study 2: Real Environment with Simulated Attacks

We Simulate Unfair Ratings using Attack Models on the Data Obtained from the Real Environment

- **Real dataset:** Real data is obtained from IMDB website. The information extracted includes userID, ratings, date, movieID, movie name, usefulness, director name, directorID, etc
- **Attack models**: The RepBad, RepSelf and the RepTrap attack models are implemented. The main goal of these attacks is to overturn the quality of the target director by providing unfair ratings

- **Detection approaches**: BRS, TRAVOS and Personalized approach
- **Simulation settings**: There are a total of 40 directors of which the first 20 are the target directors for the attackers
- **Evaluation metric**: The second robustness metric which is the **number of unfair ratings** needed by attackers to change the reputation of the target director

Case Study 2: Experimental Results

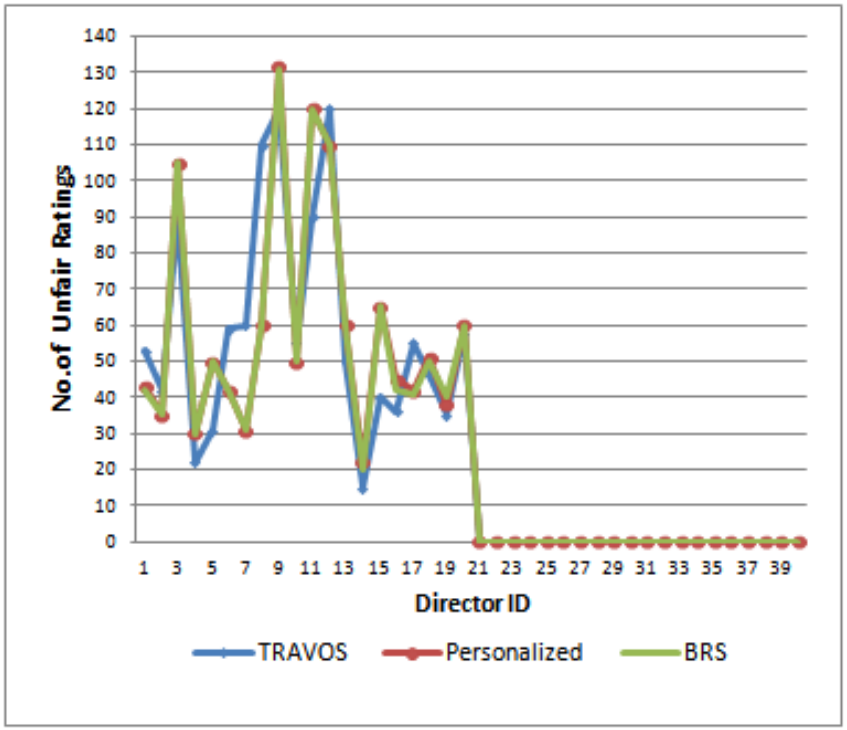

Figure: Number of Unfair Ratings Needed by RepBad.

- The average number of unfair ratings required for a successful RepBad attack is 60 for BRS, TRAVOS and Personalized
- BRS, TRAVOS, Personalized are equally robust against RepBad

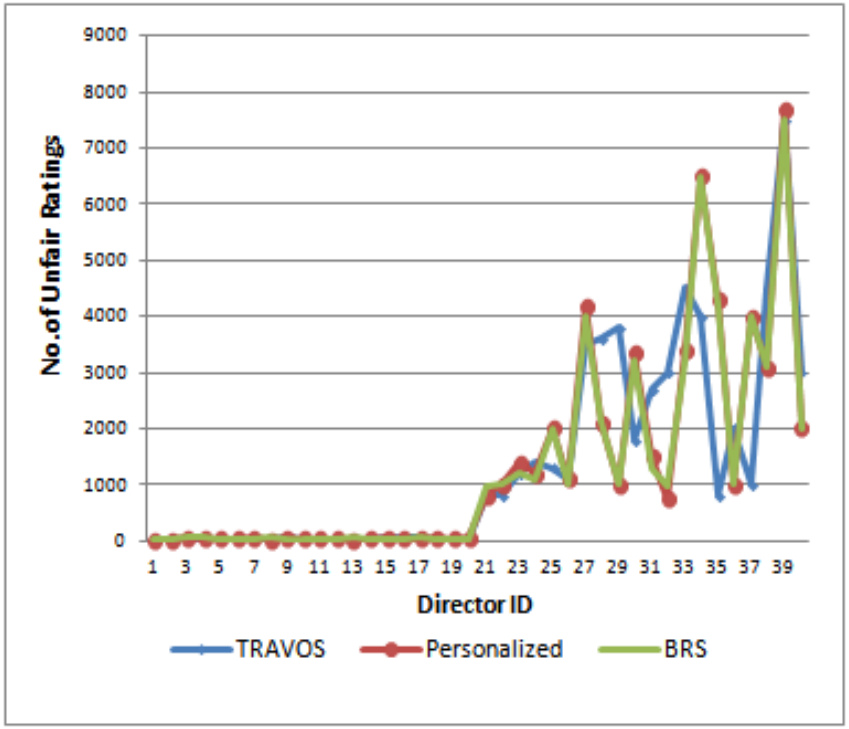

Figure: Number of Unfair Ratings Needed by RepSelf.

- The average number of unfair ratings needed to change a director's reputation is 2654, 2662, and 2607 for TRAVOS, Personalized and BRS, respectively
- The Personalized approach is found to be more robust against RepSelf attack with a maximum of 7700 unfair ratings

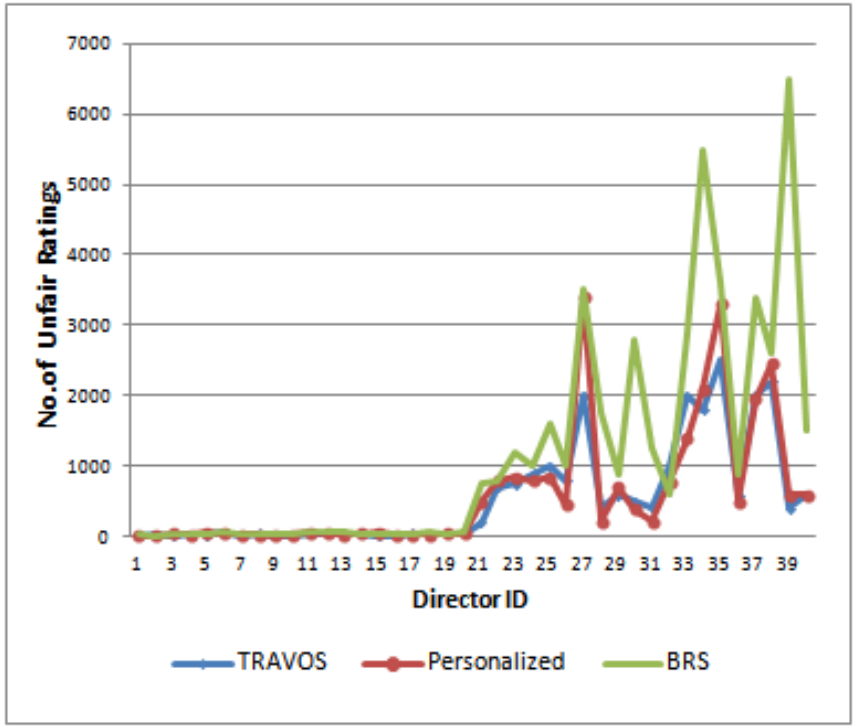

Figure: Number of Unfair Ratings Needed by RepTrap.

- The average number of unfair ratings required for a successful attack is 1102, 1182 and 2246 for TRAVOS, Personalized and BRS, respectively
- BRS is more robust than TRAVOS and Personalized against the RepTrap attack

Conclusion

- We develop a comprehensive testbed to evaluate the robustness and effectiveness of reputation systems.
- The testbed supports 3 different kinds of environments which makes it highly flexible for experimentation in a variety of settings.
- We introduce 2 novel robustness metrics to evaluate the detection approaches
- The testbed would be highly beneficial for the researchers in the field to analyze and compare their approaches with the purpose of improving their performance.

Current and Future Work

- Extend and enhance the testbed
- Detecting unfair ratings in an unknown real environment.
- Detecting unfair ratings in a multi-criteria scenario.

Publications

- "Towards a Comprehensive Testbed to Evaluate the Robustness of Reputation Systems against Unfair Rating Attacks",Athirai Aravazhi Irissappane, Siwei Jiang, Jie Zhang, Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization, Montreal, Canada, 2012
- "A Context-Aware Framework for Detecting Unfair Ratings in an Unknown Real Environment", Cheng Wan, Jie Zhang, Athirai Aravazhi Irissappane, IEEE/WIC/ACM International Conference on Web Intelligence (WI),2012.

Thank You

Questions ?