Attacking Black-box Recommendations via Copying Cross-domain User Profiles

Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma,

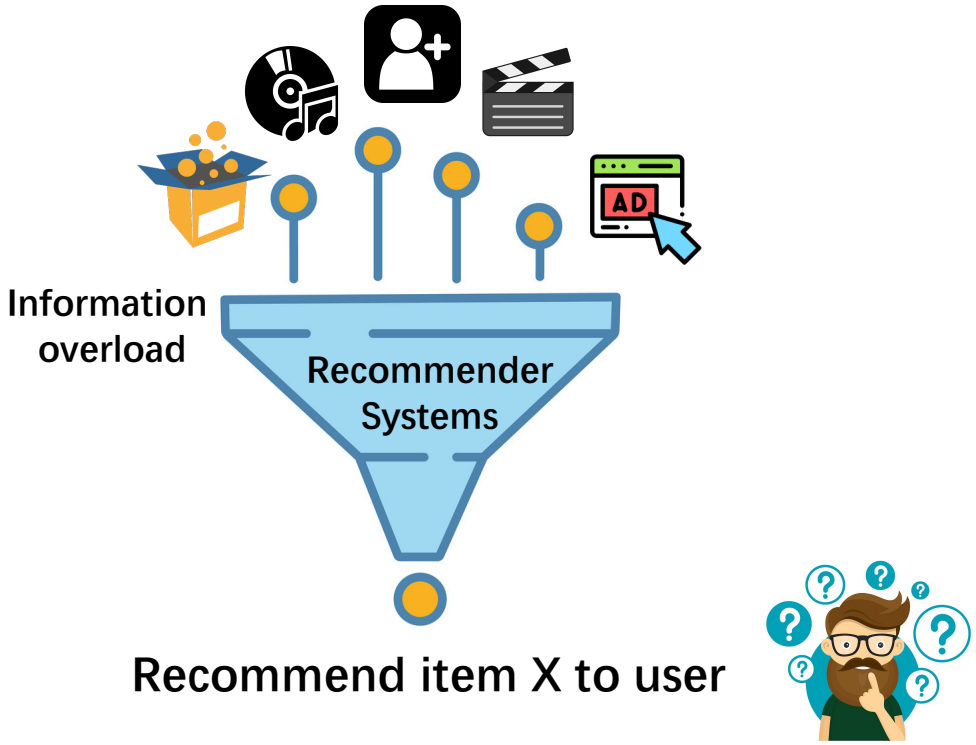
Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li

19-22 April 2021

37th IEEE International Conference on Data Engineering (ICDE)

Recommender systems

- Goal: suggest items that best fit users' preferences



amazon

淘宝网 Taobao.com

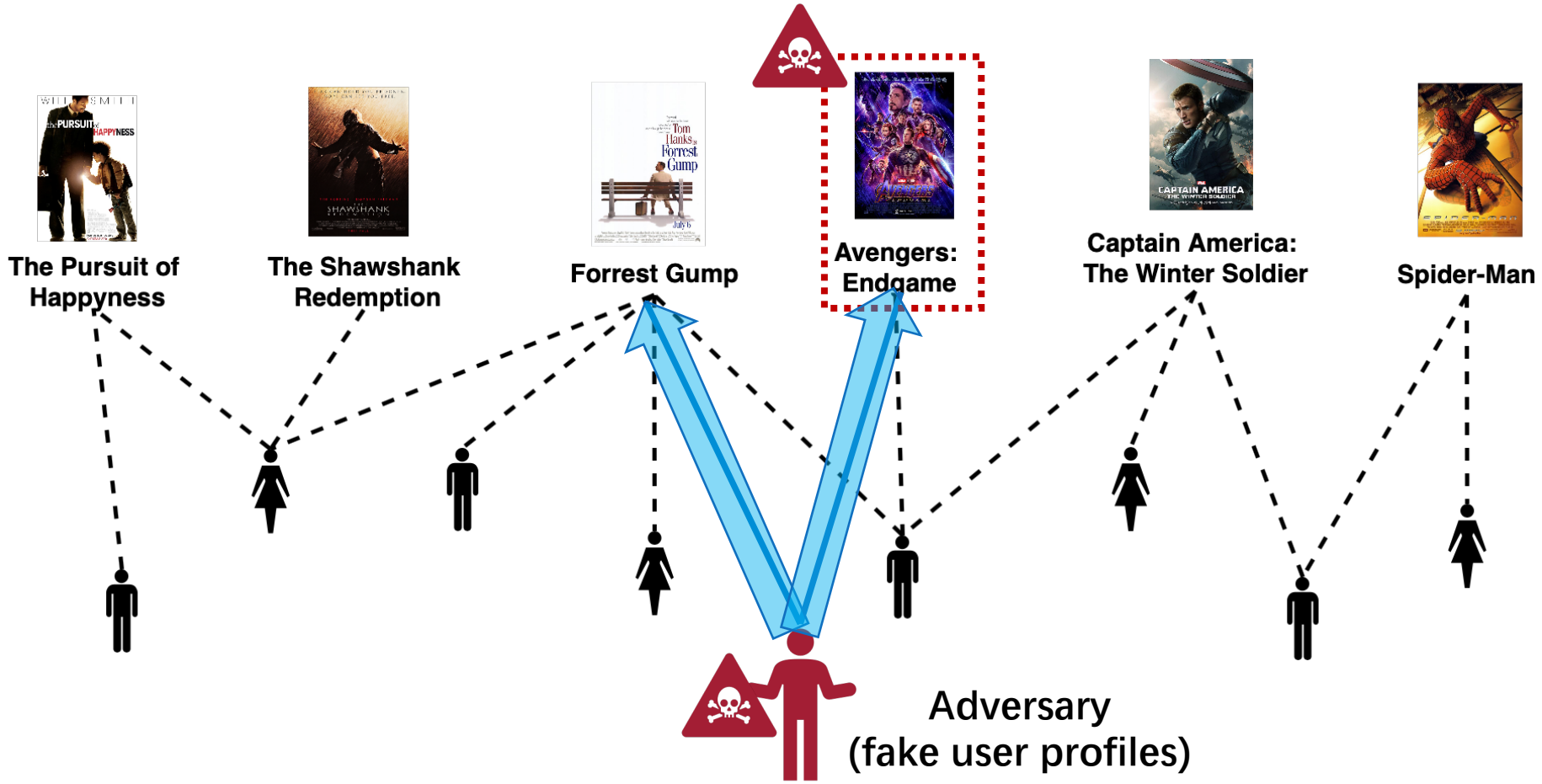
JD.COM

LinkedIn

facebook

Recommender systems

- Security (Attacking) in Recommender Systems
  - **Data poisoning attacks**: promote/demote a set of items



Attacking in recommender systems

- **Challenges in** existing attacking methods:
  - **Less "realistic" user profiles (easily detected)**

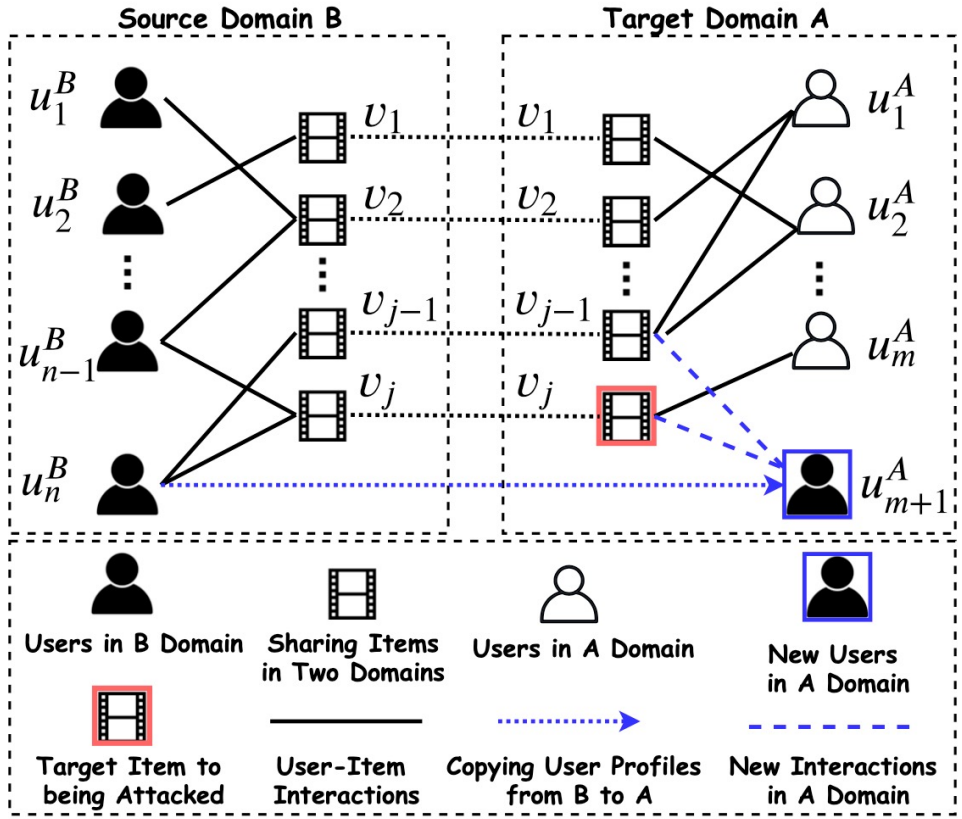
Attacking in recommender systems

- **Cross-domain Information**
  - Share a lot of items
  - Users from these platforms with **similar functionalities** also share similar behavior patterns/preferences.

Taobao JD.com

Attacking in recommender systems

- **Challenges in** existing attacking methods:
  - **Less "realistic" user profiles (easily detected)**
  - **Copy cross-domain users with real profiles from other domains**



Attacking in recommender systems

- **Challenges in existing attacking methods:**
  - **Less "realistic" user profiles (easily detected)**
    - Cross-domain Information

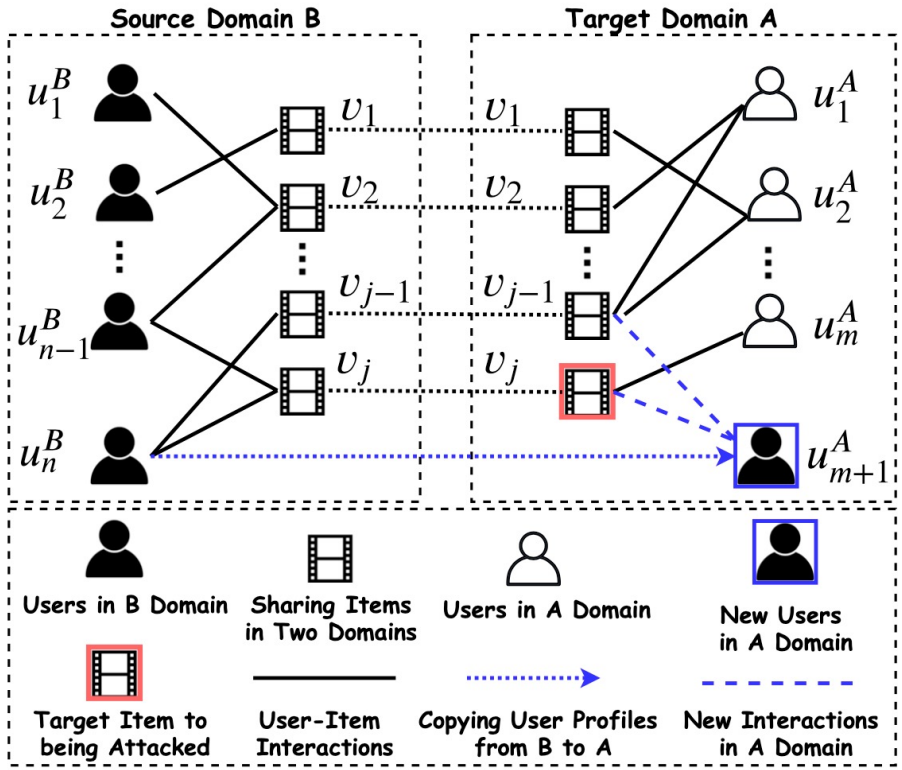
  - **White-box setting** (i.e., model architecture and parameters, and datasets)
    → impossible and unrealistic (**privacy and security**)

  - **Black-box setting**
    - ➢ Reinforcement Learning (RL) -- Query Feedback (Reward)

CopyAttack

- **Problem Statement**
  - Target RecSys A Users: $\mathcal{U}^A = \{u_1^A, u_2^A, ..., u_{n^A}^A\}$ Items: $\mathcal{V}^A = \{v_1, v_2, ..., v_{m^A}\}$
    - User profile: $P_{u_i}^A = \{v_1 \rightarrow ... \rightarrow v_j \rightarrow ... \rightarrow v_l\}$
    - Item profile: $P_{v_j}^A = \{u_1^A \rightarrow ... \rightarrow u_i^A \rightarrow ... \rightarrow u_o^A\}$
  - Source RecSys B: Users: $\mathcal{U}^B$ Items: $\mathcal{V}^B$
    - User Profile: $\mathcal{P}_{\mathcal{U}}^B = \{v_1 \rightarrow ... \rightarrow v_j \rightarrow ... \rightarrow v_l\}$
  - Overlapping items: $\mathcal{V} = \mathcal{V}^A \cap \mathcal{V}^B$
- Goal: $\mathcal{U}^{A\prime} = \mathcal{U}^A \cup \mathcal{U}^{B \rightarrow A}$

$$y_{i,>k}^A = \{v_{[1]}, v_{[2]}, ..., v_{[k]}\} = Rec(P_{u_i}^A, \mathcal{P}_{\mathcal{V}}^A)$$



CopyAttack

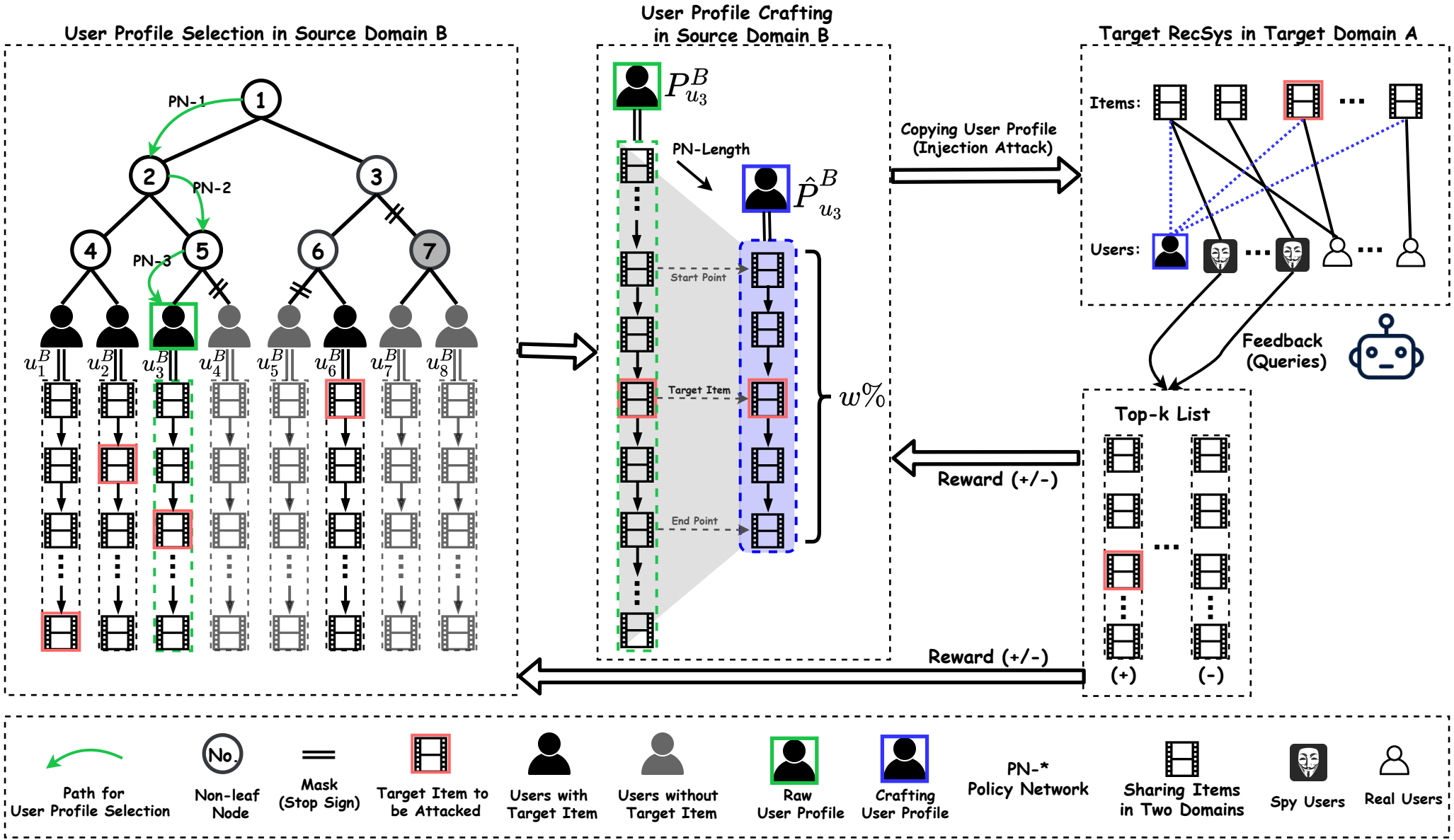
- **Attacking RL Environment**
  - Action A: user profiles in source domain B
  - Reward R (Hitting Ratio, HR):
    - **Spy users**

$$r(s_t, a_t) = \frac{1}{|\mathcal{U}_*^A|} \sum_{i=1}^{|\mathcal{U}_*^A|} HR(u_{i*}^A, v_*, k)$$

$$HR(u_{i*}^A, v_*, k) = \begin{cases} 1, & v_* \in y_{u^*,>k}, \\ 0, & v_* \notin y_{u^*,>k} \end{cases}$$

  - Terminal: reach the budget or successfully satisfy the promotion task

CopyAttack


User Profile Selection in Source Domain B
User Profile Crafting in Source Domain B
Target RecSys in Target Domain A
PN-1
PN-2
PN-3
PN-Length
Start Point
Target Item
End Point
$w\%$
Copying User Profile (Injection Attack)
Items:
Users:
Feedback (Queries)
Top-k List
(+)
(-)
Reward (+/-)
Reward (+/-)
Path for User Profile Selection
Non-leaf Node
Mask (Stop Sign)
Target Item to be Attacked
Users with Target Item
Users without Target Item
Raw User Profile
Crafting User Profile
PN-* Policy Network
Sharing Items in Two Domains
Spy Users
Real Users

CopyAttack

- User Profile Selection
  - Construct hierarchical clustering tree
  - Masking Mechanism - specific target items
  - Hierarchical-structure Policy Gradient

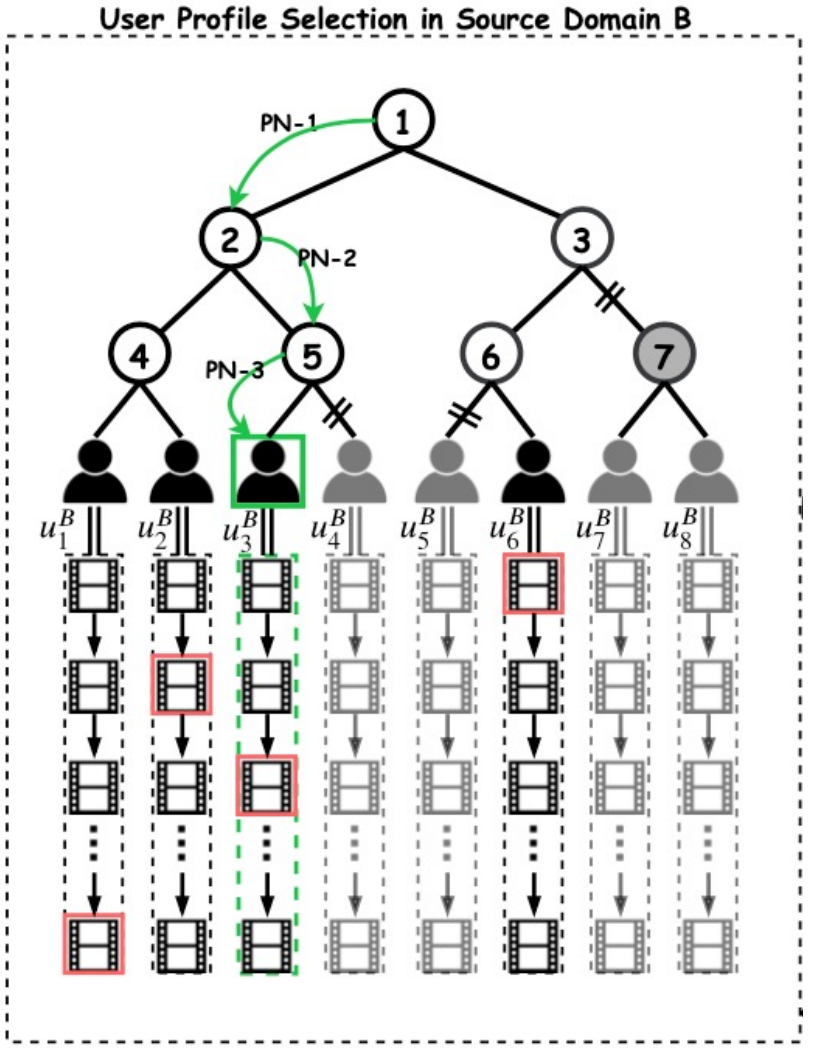
$$a_t^u = \left\{a_{[t,1]}^u, a_{[t,2]}^u, ..., a_{[t,d]}^u\right\}$$

$$p^u(a_t^u|s_t^u) = \prod_d^d p_d^u(a_t^u|\cdot, s_t^u)$$
$$= p_d^u(a_{[t,d]}^u|s_t^u) \cdot p_{d-1}^u(a_{[t,d-1]}^u|s_t^u) \cdots p_1^u(a_{[t,1]}^u|s_t^u)$$

$$\mathbf{x}_{v_*} = RNN(\mathcal{U}_t^{B\to A}),$$

$$p_i^u(\cdot|s_t^u) = softmax(MLP([\mathbf{q}_{v_*}^B \oplus \mathbf{x}_{v_*}]|\theta_i^u))$$

Time Complexity： $\mathcal{O}(|\mathcal{U}^B|) \longrightarrow \mathcal{O}(d \times |\mathcal{U}^B|^{1/d})$



CopyAttack

- User Profile Crafting
  - Clipping operation to craft the raw user profiles

$$W = \{10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%\}$$

  - Sequential patterns (forward/backward)

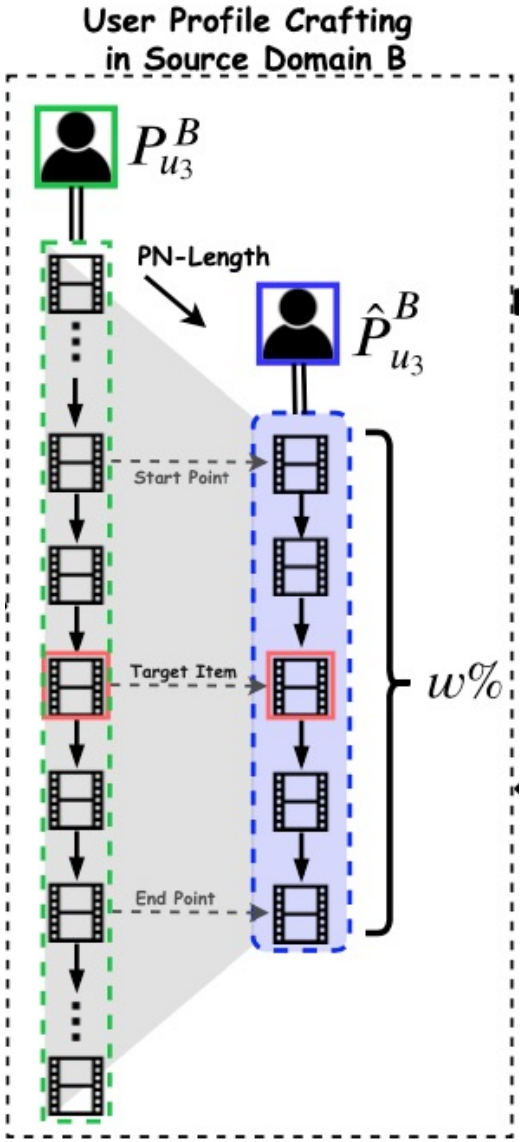
w = 50%

Example:

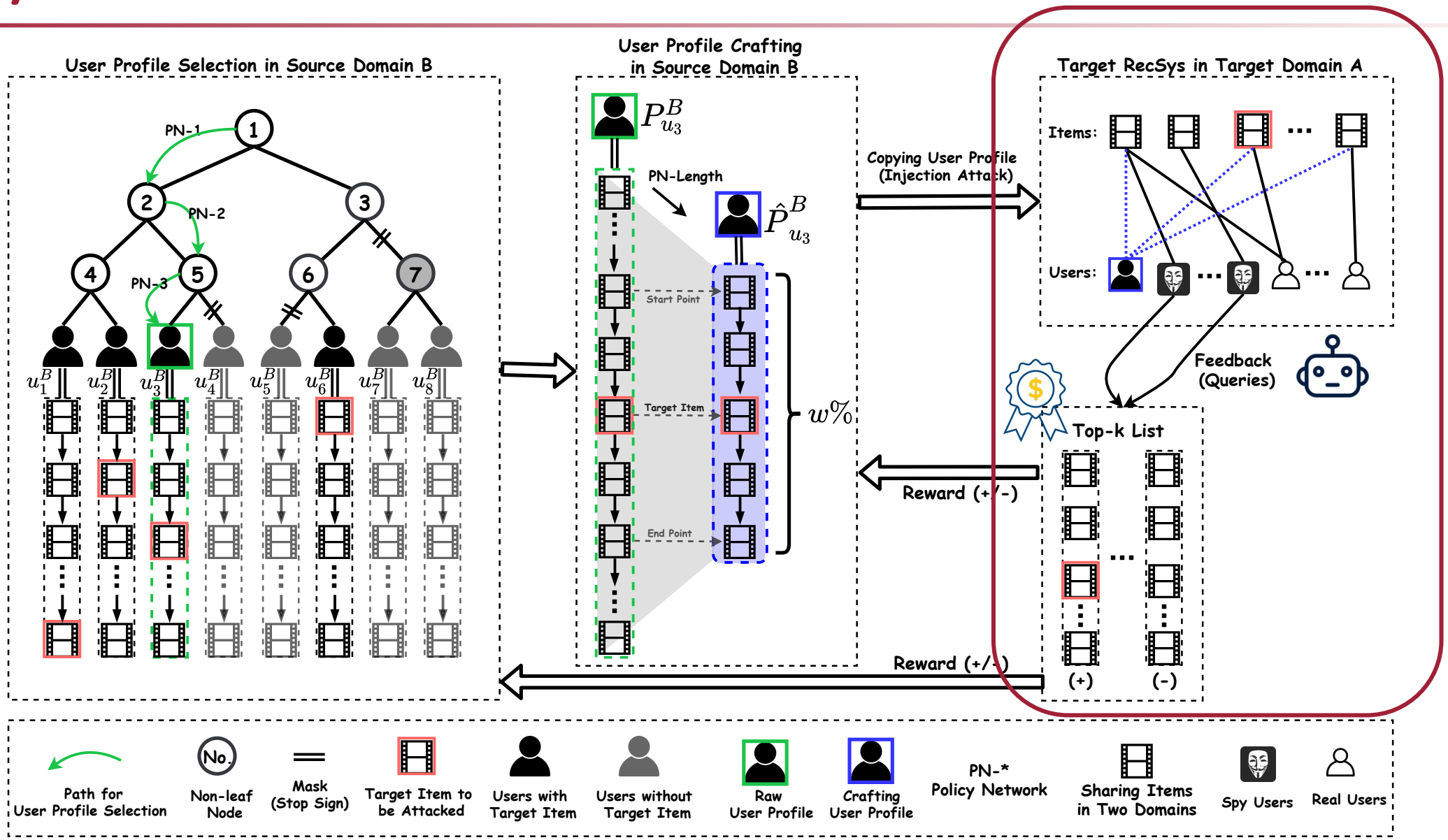
$$P_{u_i}^B = \{v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \rightarrow v_{5*} \rightarrow v_6 \rightarrow v_7 \rightarrow v_8 \rightarrow v_9 \rightarrow v_{10}\}$$

$$\hat{P}_{u_i}^B = \{v_3 \rightarrow v_4 \rightarrow v_{5*} \rightarrow v_6 \rightarrow v_7\}$$

$$p^l(\cdot|s_t^l) = softmax(MLP([\mathbf{p}_i^B \oplus \mathbf{q}_{v_*}^B]|\theta^l))$$



CopyAttack


User Profile Selection in Source Domain B
PN-1
PN-2
PN-3
1
2
3
4
5
6
7
u_1^B u_2^B u_3^B u_4^B u_5^B u_6^B u_7^B u_8^B
User Profile Crafting in Source Domain B
$P_{u_3}^B$
PN-Length
$\hat{P}_{u_3}^B$
Start Point
Target Item
End Point
$w\%$
Copying User Profile (Injection Attack)
Target RecSys in Target Domain A
Items:
Users:
Feedback (Queries)
Top-k List
(+)
(-)
Reward (+/-)
Reward (+/-)
Path for User Profile Selection
No.
Non-leaf Node
Mask (Stop Sign)
Target Item to be Attacked
Users with Target Item
Users without Target Item
Raw User Profile
Crafting User Profile
PN-* Policy Network
Sharing Items in Two Domains
Spy Users
Real Users

Thank You

Wenqi Fan, wenqifan03@gmail.com

Please see my homepage for more details:
https://wenqifan03.github.io

Attacking Black-box Recommendations via Copying Cross-domain User Profiles

Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma,

Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li

19-22 April 2021

37th IEEE International Conference on Data Engineering (ICDE)