PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems

Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum
Max Planck Institute for Informatics, Saarbruecken, Germany
PRINCE: **P**rovider-side **I**nterpretability with **C**ounterfactual **E**xplanations in Recommender Systems

aka

On the Feasibility and Usefulness of Action-based Explanations in Recommenders

Azin Ghazimatin, Oana Balalau, **Rishiraj Saha Roy**, and Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany
Problem: Why did I receive this recommendation?
Problem: Why did I receive this recommendation?
Problem Setup

Nodes: Users, Items, Categories
Edges: Actions or Similarity, Directed, Weighted
Problem Setup

Framework: Heterogeneous Information Network

Nodes: Users, Items, Categories
Edges: Actions or Similarity, Directed, Weighted
Problem Setup

**Framework:** Heterogeneous Information Network

**Recommender Model:** Known

**Nodes:** Users, Items, Categories

**Edges:** Actions or Similarity, Directed, Weighted
Problem Setup

**Framework:** Heterogeneous Information Network

**Recommender Model:** Known

**In this work:** RecWalk (WSDM 2019)

**Nodes:** Users, Items, Categories

**Edges:** Actions or Similarity, Directed, Weighted
Problem Setup

**Framework:** Heterogeneous Information Network

**Recommender Model:** Known

**In this work:** RecWalk (WSDM 2019)

**Characterized by:** Transition Matrix W, Rec. Score PPR(u, rec)

**Nodes:** Users, Items, Categories

**Edges:** Actions or Similarity, Directed, Weighted
Existing approaches based on paths
Concerns with paths: Too many
Concerns with paths: Potential privacy breaches
Concerns with paths: Not actionable

Not clear if acting upon these edges changes rec

Users have no control on this part of the graph
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on explanations

**Actionable:** User’s own actions
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on explanations

**Concise:** Use minimal sets
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on recommender explanations
PRINCE: A fresh perspective on recommender explanations

Scrubtable: Use counterfactual setup
PRINCE: A fresh perspective on recommender explanations

★ **Actionable**: Ground explanations in user’s own actions
★ **Concise**: Use minimal sets of actions for presentation
★ **Scrubtable**: Use a counterfactual setup for derivation

**User u:** Why did I receive this recommendation “Jack Wolfskin backpack”?

**PRINCE:** You **bought** “Adidas Hiking Shoes”;  
You **reviewed** “Nikon Coolpix Camera” with “Sleek! Handy on hikes!”;  
You **rated** “Intenso Travel Power Bank” highly.

If you **had not** done these actions:  
“iPad Air” **would have replaced** “Jack Wolfskin backpack”.
Formal problem statement for counterfactual explanation

Find **minimal** set of **actions** whose **removal** displaces the top recommended item
Naïve algorithm: Exponential time complexity!!

Try all action subsets whose removal displaces rec and find minimum

In time $2^{|\text{Actions}|}$
PRINCE has polynomial time complexity: How?
PRINCE iterates over items and not action subsets
PRINCE iterates over items and not action subsets
PRINCE iterates over items and not action subsets: Item $i_1$
PRINCE iterates over items and not action subsets: Item $i_2$
PRINCE iterates over items and not action subsets: Item $i_3$ ...
PRINCE iterates over items and not action subsets: Item $i_3$ ...

- **Iterate** over items to find actions for **swapping** rec and rec*

- **Sort** actions by contribution within each iteration

- **Report** minimal set

  $O(\ |\text{Items}| \times \ |\text{Actions}| \times \log |\text{Actions}|)$
Core of PRINCE: How do we swap orders of rec and rec*?
Swapping items: Actions connect to user $u$ to neighbors $n_i$
Swapping items: Define contributions of neighbors $n_i$ to rec

PPR allows defining **contributions** of neighbors $n_i$ to rec!

PPR($u$, rec) **weighted average** over PPR($n_i$, rec)

Contribution of $n_i$ to rec: $W(u, n_i) \cdot PPR(n_i, rec)$
Swapping items: Define contributions of neighbors $n_i$ to $\text{rec}^*$

Contrib. of $n_i$ to $\text{rec}^*$: $W(u, n_i) \cdot PPR(n_i, \text{rec}^*)$
Sort neighbors by highest contribution to rec and least to rec*

Difference in contributions: \( W(u, n_i) \cdot PPR(n_i, rec) - PPR(n_i, rec^*) \)
Delete neighbors greedily until replacement

\[ W(u, n_i) \cdot PPR(n_i, rec) - PPR(n_i, rec^*) \]
Delete neighbors greedily until replacement

\[ W(u, n_i) \cdot PPR(n_i, \text{rec}) - PPR(n_i, \text{rec}^*) \]
Delete neighbors greedily until replacement

Difference in contributions: \( W(u, n_i) \cdot PPR(n_i, rec) - PPR(n_i, rec^*) \)
Why is this greedy strategy optimal? Theorems!

★ PPR score for rec when action subset $A^*$ is removed from the graph

can be computed as product of two components, and only one matters

$$PPR(u, rec \mid A^*) = \frac{1 - \alpha}{\alpha} \cdot PPR(u, u \mid A^*) \cdot \sum_{(u,n_i)\in A\setminus A^*} W(u, n_i \mid A^*) \cdot PPR(n_i, rec \mid A)$$

★ The condition for replacement can be reformulated as

$$PPR(u, rec \mid A^*) < PPR(u, rec^* \mid A^*) \Leftrightarrow \sum_{(u,n_i)\in A\setminus A^*} W(u, n_i) \cdot [PPR(n_i, rec \mid A) - PPR(n_i, rec^* \mid A)] < 0$$
Putting it all together: The PRINCE Algorithm

**Input:** \( G = (V, E, \theta), u, rec, I \)

**Output:** \( A^*, rec^* \)  
Counterfactual explanation

**Initialize:** \( A^* \leftarrow A, rec^* \leftarrow rec \)

**for each** \( i \in I \)  
Iterate over each item \( i \)

\[ A^i = SwapOrder(G, u, rec, i) \]  
Core: Find actions to swap \( rec \) and \( i \)

**if** \( |A^i| < |A^*| \)  
Minimality check

\[ A^* \leftarrow A^i, rec^* \leftarrow i \]

**return** \( A^*, rec^* \)  
Report minimal set

\[ return \]
Experimental setup

★ **Datasets**  Amazon, Goodreads

★ **Graph size**  2k users, 50k items, 58k actions (ratings + reviews), 40 categories (for Amazon)

★ **Replacement item**  From top-k recommendations

★ **Evaluations**  Graph measurements, User study
Graph results

- **Baselines:** HC (Highest Contributions), SP (Shortest Paths)
- **Metric:** Explanation size
- PRINCE is **difficult to approximate**
- Explanations **shrink** with increasing $k$
- PRINCE is **efficient**

<table>
<thead>
<tr>
<th>k</th>
<th>PRINCE</th>
<th>HC</th>
<th>SP</th>
<th>PRINCE</th>
<th>HC</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5.09*</td>
<td>6.87</td>
<td>7.57</td>
<td>2.05*</td>
<td>2.86</td>
<td>5.38</td>
</tr>
<tr>
<td>5</td>
<td>3.41*</td>
<td>4.62</td>
<td>5.01</td>
<td>1.66*</td>
<td>2.19</td>
<td>4.37</td>
</tr>
<tr>
<td>10</td>
<td>2.66*</td>
<td>3.66</td>
<td>4.15</td>
<td>1.43</td>
<td>1.45</td>
<td>3.28</td>
</tr>
<tr>
<td>15</td>
<td>2.13*</td>
<td>3.00</td>
<td>3.68</td>
<td>1.11</td>
<td>1.12</td>
<td>2.90</td>
</tr>
<tr>
<td>20</td>
<td>1.80*</td>
<td>2.39</td>
<td>3.28</td>
<td>1.11</td>
<td>1.12</td>
<td>2.90</td>
</tr>
</tbody>
</table>

Runtimes at $k = 5$:
- 1.3 milliseconds (Precomputed)
- 121.6 seconds (Online)
Mechanical Turk study on usefulness

- Explanations for 200 (user, rec) pairs from Amazon judged for **usefulness**
- On a scale of **1 – 3** (most useful)
- **Baseline**: CredPaths (Yang et al. ICDM 2018)
- **Three Master** Workers per HIT
- **Honeypots** in HITs
- Controlled for **presentation, size and comparison** biases

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>#Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRINCE</td>
<td>1.91*</td>
<td>0.66</td>
<td>200</td>
</tr>
<tr>
<td>CredPaths</td>
<td>1.78*</td>
<td>0.63</td>
<td>200</td>
</tr>
<tr>
<td>PRINCE (Size = 1)</td>
<td>1.87*</td>
<td>0.66</td>
<td>154</td>
</tr>
<tr>
<td>PRINCE (Size = 2)</td>
<td>1.88*</td>
<td>0.70</td>
<td>28</td>
</tr>
<tr>
<td>PRINCE (Size = 2)</td>
<td>2.21*</td>
<td>0.52</td>
<td>18</td>
</tr>
</tbody>
</table>

PRINCE is judged **more useful** at all explanation sizes!

Qualitative survey on usefulness in paper!
Contributions in PRINCE

★ Counterfactual evidence for discovering causal explanations in heterogeneous information networks

★ Explanations in terms of users’ own actions – feasible and useful!!

★ Optimal algorithm explores action space to find minimal subsets

★ Helps provider present actionable, concise, and scrutable explanations!!

Code  github.com/azinmatin/prince/