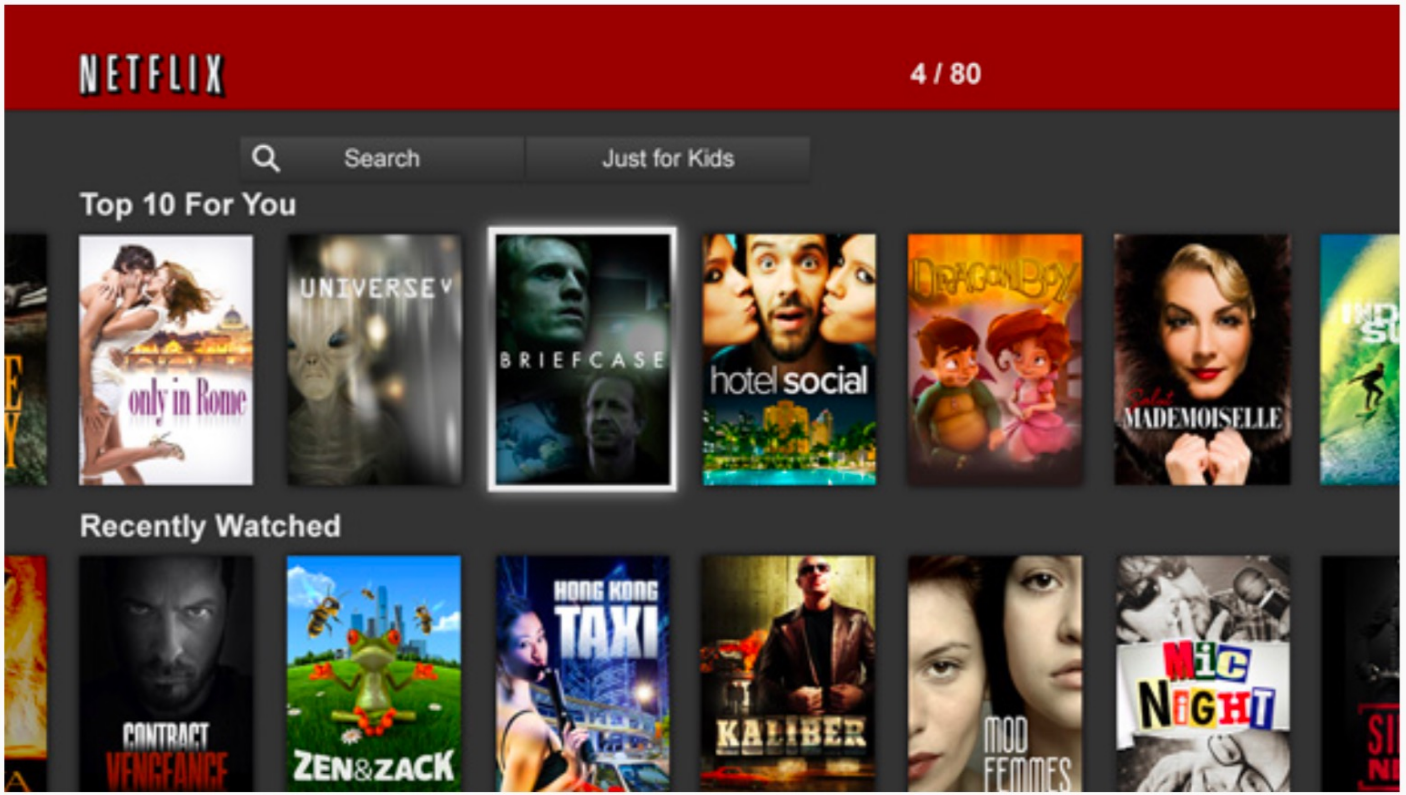
inventors for the digital world

Collaborative Filtering Under a Sybil Attack: Similarity Metrics do Matter!

Davide Frey,
joint work with: Antoine Boutet, Florestan De Moor, Rachid Guerraoui, Anne-Marie Kermarrec and Antoine Rault

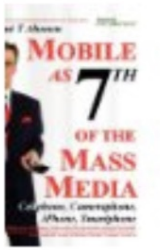
Recommender Systems



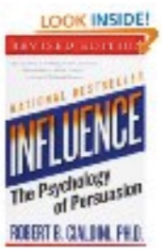
Netflix: 75% of views driven by recommendation

Recommender Systems

Customers Who Bought This Item Also Bought



Mobile as 7th of the Mass Media: Cellphone, …
Tomi Ahonen
(3)
Hardcover
$44.99

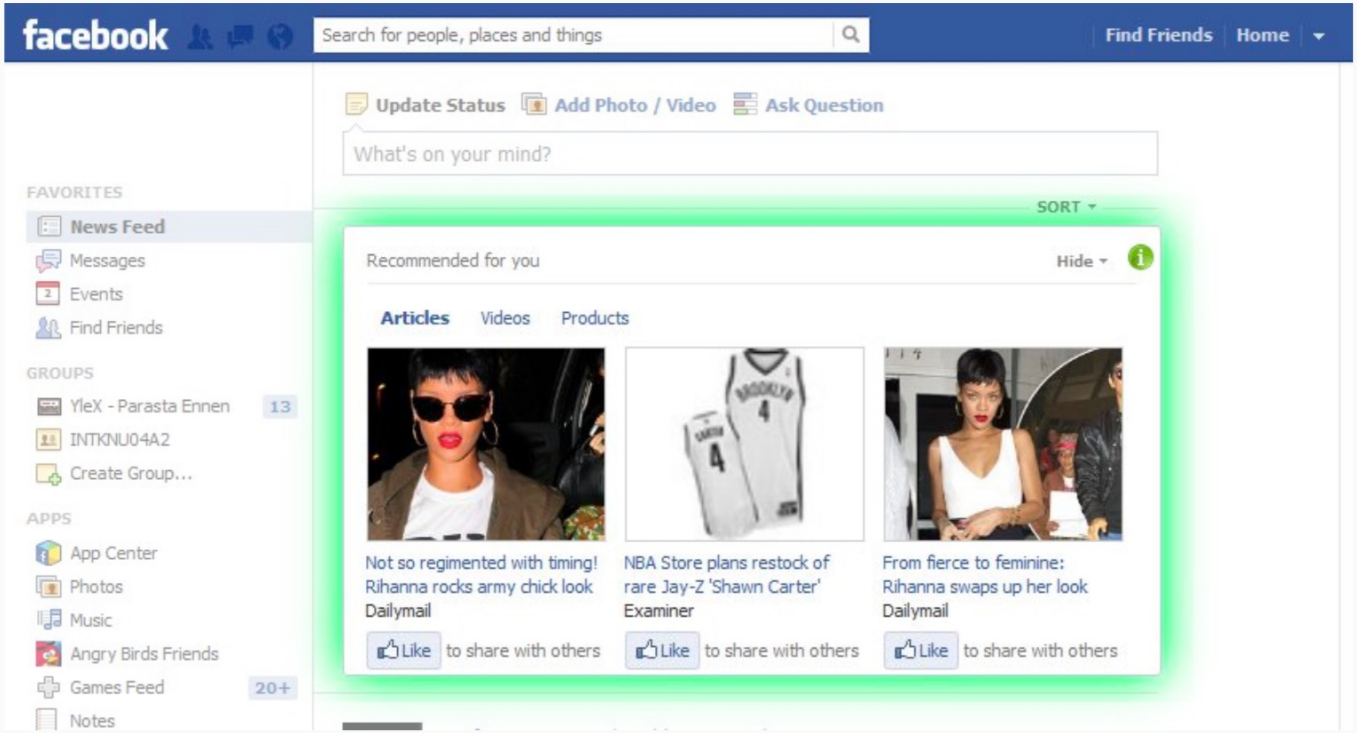


Influence: The Psychology of …
› Robert B. Cialdini
(495)
Paperback
$13.46

Digital Korea: Convergence of …
Tomi Ahonen
(5)
Hardcover
$44.96

Amazon: +29% sales from recommendation

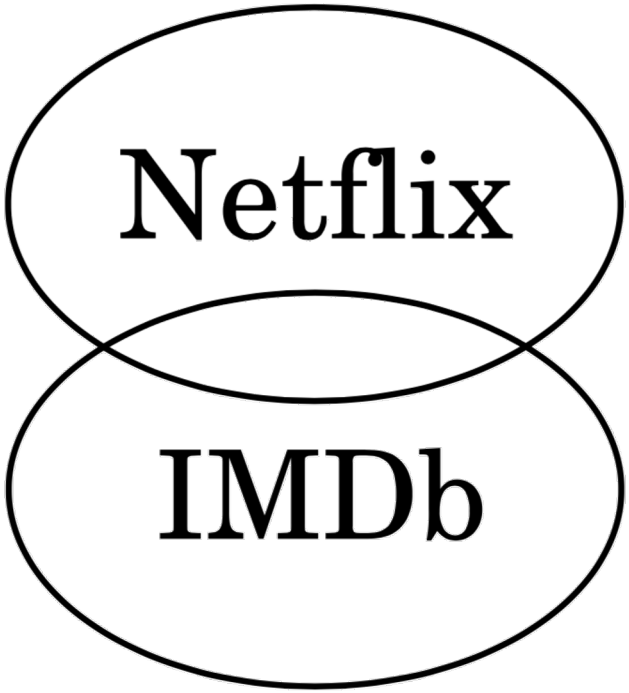
Recommender Systems



Facebook: News Feed is like a RS

Recommender and Privacy

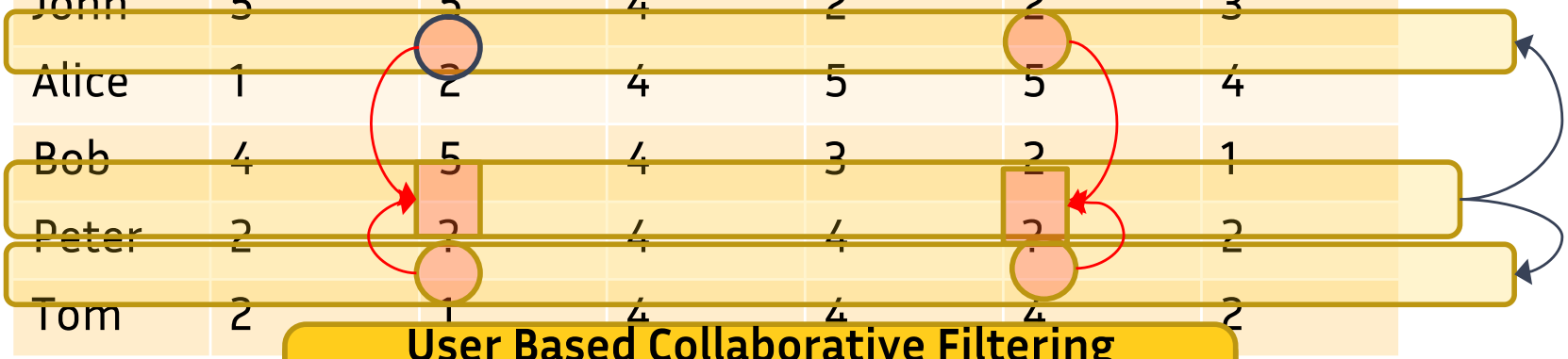
Example: Netflix Prize De-anonymization



Collaborative Filtering

- **Each user associated with a « user profile »**
  - List of items and associated ratings

	Avengers	Iron Man	Star Wars	The Godfather	Forrest Gump	Die Hard
John	5	5	4	2	2	3
Alice	1	2	4	5	5	4
Bob	4	5	4	3	2	1
Peter	2	?	4	4	?	2
Tom	2	1	4	4	4	2



User Based Collaborative Filtering
Identify the K-nearest neighbors of each user
The KNN provide recommendations

Attacker Model

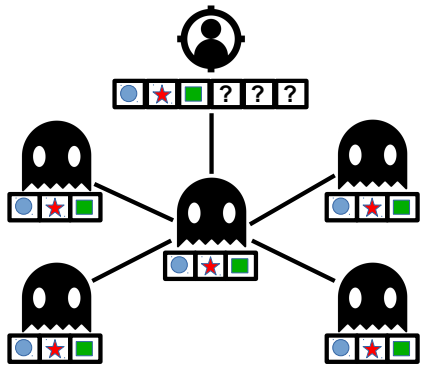
- **Adversary targeting one user** [CKNFS11]
  - Access to auxiliary information
    - A fraction p of the target user's profile
      - i.e. $p|t|$ items, t being the target profile
    - Obtained through public or insider information
  - Ability to create a number of fake users (Sybils)
  - Knowledge of k in KNN

[CKNFS11] proposed the attack but did not evaluate it.

[CKNFS11] Calandrino, Kilzer, Narayanan, Felten, Shmatikov; "You Might Also Like:" Privacy Risks of Collaborative Filtering; S&P 2011

Attack in Practice

- **Isolate the target user**
  - Create k fake identities (Sybils) using the auxiliary information
  - Success: obtain a neighborhood consisting of k-1 Sybils and the target.



- **Use recommendations to guess items**
  - Sybils obtain recommendations
  - They pool them together
  - If success (above) is met, then all non Sybil items come from the target.

Our Contributions

- Attack evaluation on State of the Art
- Novel attack resilient metric
- Evaluation on standard and improved attack.

Experimental Protocol

- Recommender system based on Apache Mahout
- Three real-world datasets

	# users	# items	# ratings	rating type
ML-1	943	1,682	100,000	$[1:5]$
Jester-1-1	24,983	100	1,810,455	$[-10.0:10.0]$
MovieTweetings	24,921	15,142	212,835	$[0:10]$

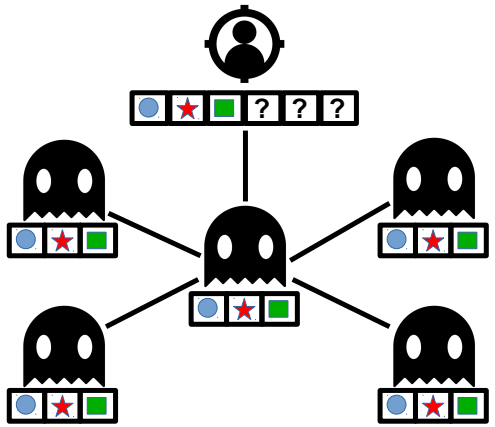
- Attack Success Metrics
- Evaluation on 7 similarity metrics

Attack Success Metrics

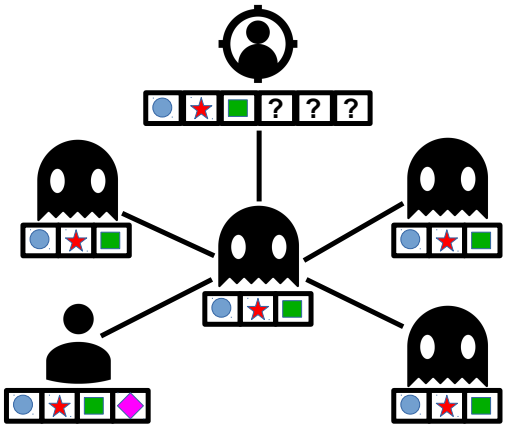
- **Expected Neighborhood**
  - Fraction of Sybils create the desired neighborhood
- **Accuracy**
  - Percentage of guesses that belong to the target's profile
- **Yield**
  - Number of guesses obtained in recommendation step (each Sybil asks for 5 recommendations)

Expected Neighborhood

The attack succeeds if Sybils construct a neighborhood consisting of K-1 other Sybils plus the target.



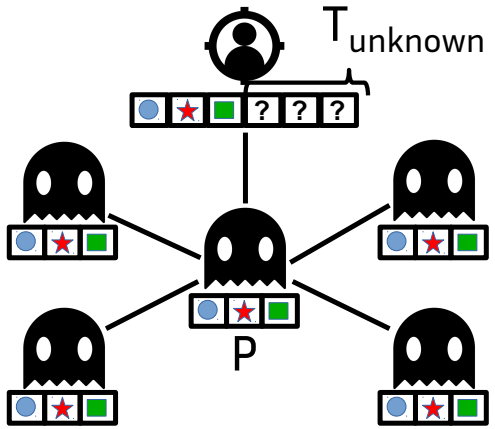
Expected Neighborhood
Attack Succeeds



Not Expected Neighborhood
Cannot Tell

Accuracy

Fraction of the recommended items that belong to the target's profile.



$$\text{Accuracy} = \frac{(\bigcup_{i=1}^{k} R_i \setminus P) \cap T_{\text{unknown}}}{(\bigcup_{i=1}^{k} R_i \setminus P)}$$

Similarity Metrics

Several similarity metrics for KNN computation

$$Cos(u,n) = \frac{r_u \cdot r_n}{||r_u|| \cdot ||r_n||}$$

$$CosineAvg(u,n) = \frac{\sum_{i \in I_u \cup I_n} \tilde{r}_{u,i} \times \tilde{r}_{n,i}}{||\tilde{r}_u|| \cdot ||\tilde{r}_n||} \quad ; \quad \tilde{r}_{x,i} = \begin{cases} r_{x,i} & \text{if } i \in I_x \\ \overline{r_x} & \text{if } i \notin I_x \end{cases}$$

$$Pearson(u,n) = \frac{\sum_{i \in I_{u,n}} (r_{u,i} - \overline{r_u})(r_{n,i} - \overline{r_n})}{\sqrt{\sum_{i \in I_{u,n}} (r_{u,i} - \overline{r_u})^2 \sum_{i \in I_{u,n}} (r_{n,i} - \overline{r_n})^2}}$$

$$WUP\text{-}u(u,n) = \frac{\sum_{i \in I_{u,n}} r_{u,i} \times r_{n,i}}{\sqrt{\sum_{i \in I_{u,n}} (r_{u,i})^2} \times \sqrt{\sum_{i \in I_n} (r_{n,i})^2}}$$

$$Cos\text{-}overlap(u,n) = \frac{r_u \cdot r_n}{\sqrt{\sum_{i \in I_{u,n}} (r_{u,i})^2} \times \sqrt{\sum_{i \in I_{u,n}} (r_{n,i})^2}}$$

$$WUP\text{-}n(u,n) = \frac{\sum_{i \in I_{u,n}} r_{u,i} \times r_{n,i}}{\sqrt{\sum_{i \in I_u} (r_{u,i})^2} \times \sqrt{\sum_{i \in I_{u,n}} (r_{n,i})^2}}$$

$$Jaccard(u,n) = \frac{|r_u \cap r_n|}{|r_u \cup r_n|}$$

Similarity Metrics

Several similarity metrics for KNN computation

- Cosine similarity
- Pearson
- Cosine average
- Jaccard
- Cosine overlap
- WUP-u
- WUP-n

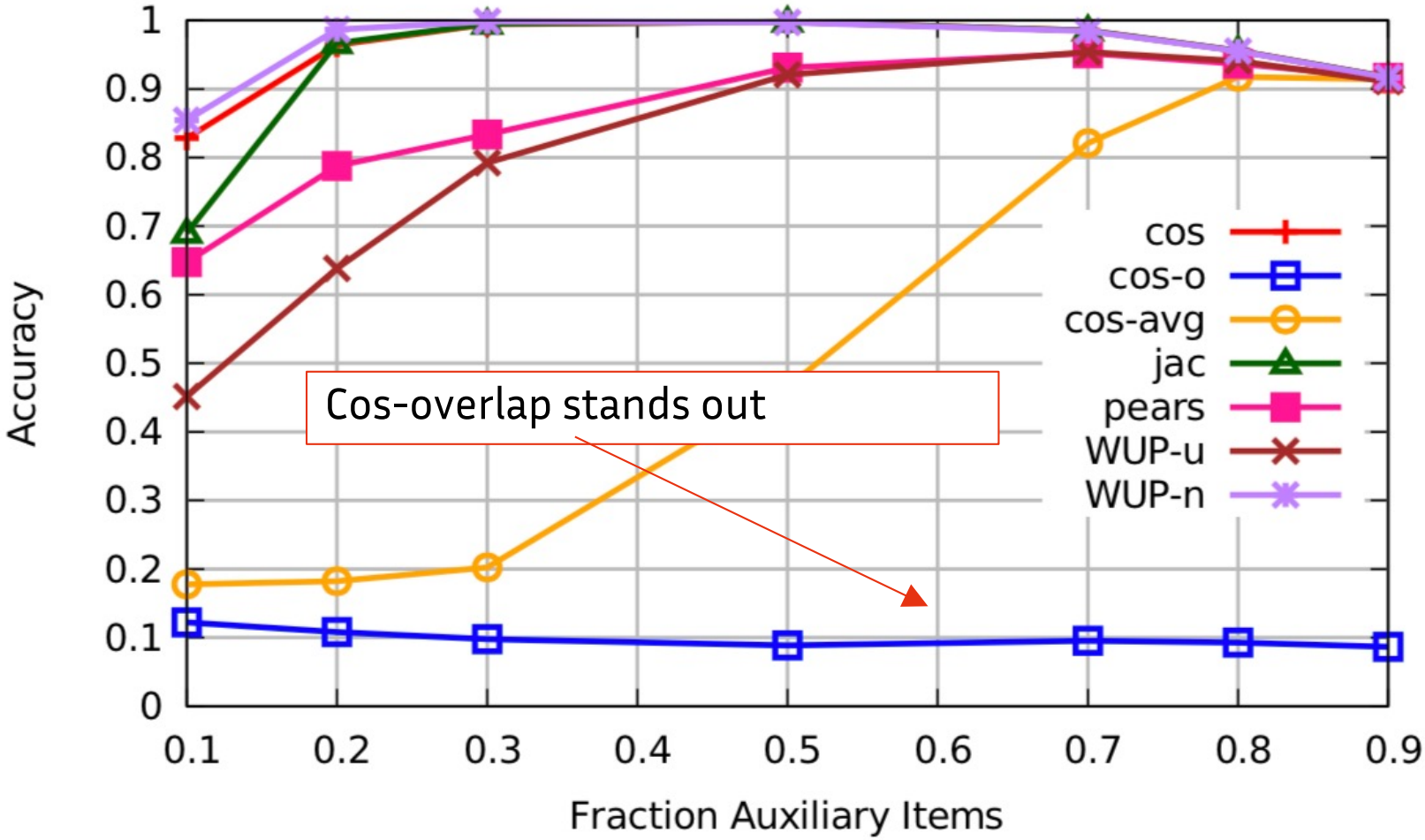
Well known

Default in Mahout

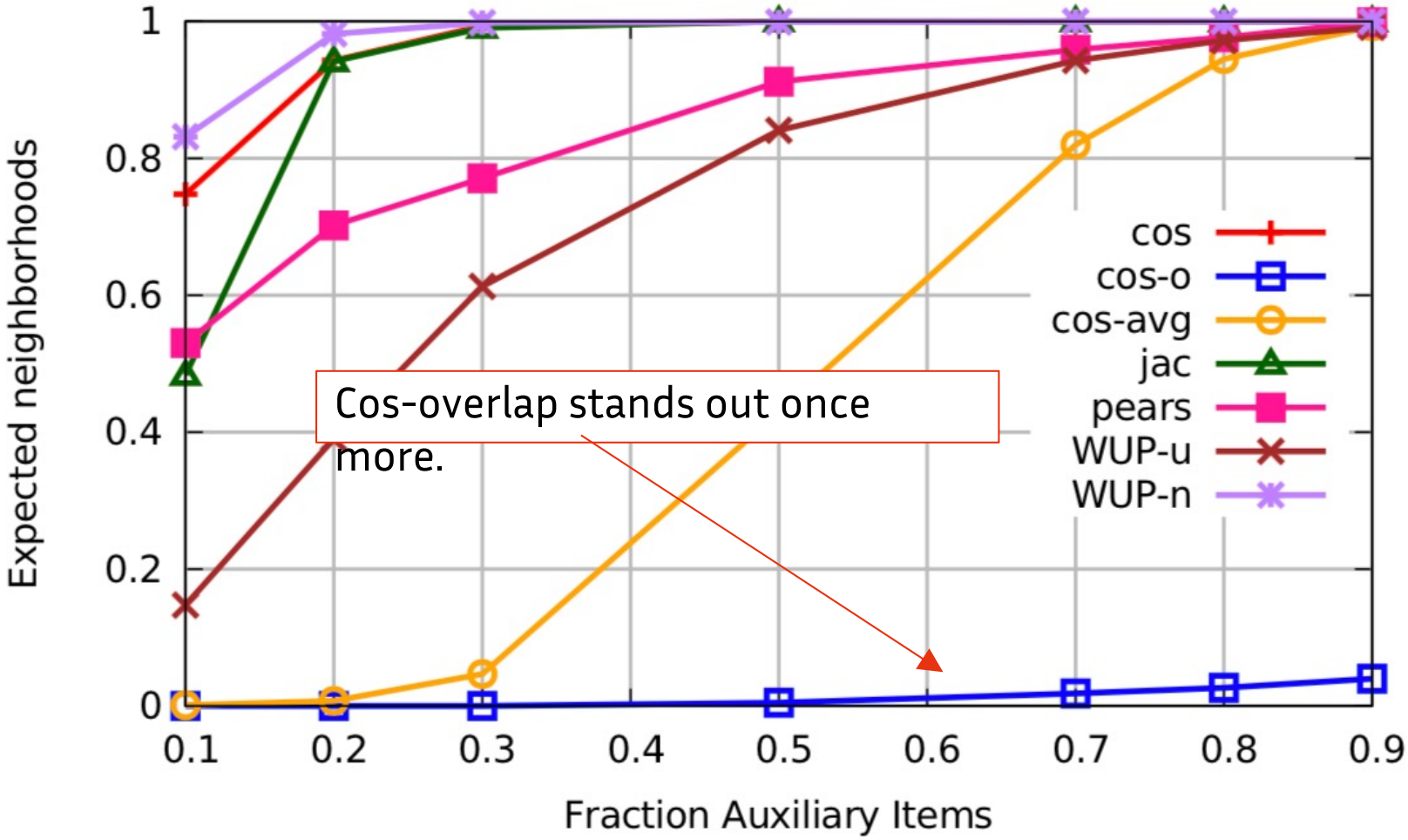
Defined in [BFGJK13]

[BFGJK13] A. Boutet, D. Frey, R. Guerraoui, A. Jégou, A. –M. Kermarrec: WHATSUP: A Decentralized Instant News Recommender. IPDPS 2013

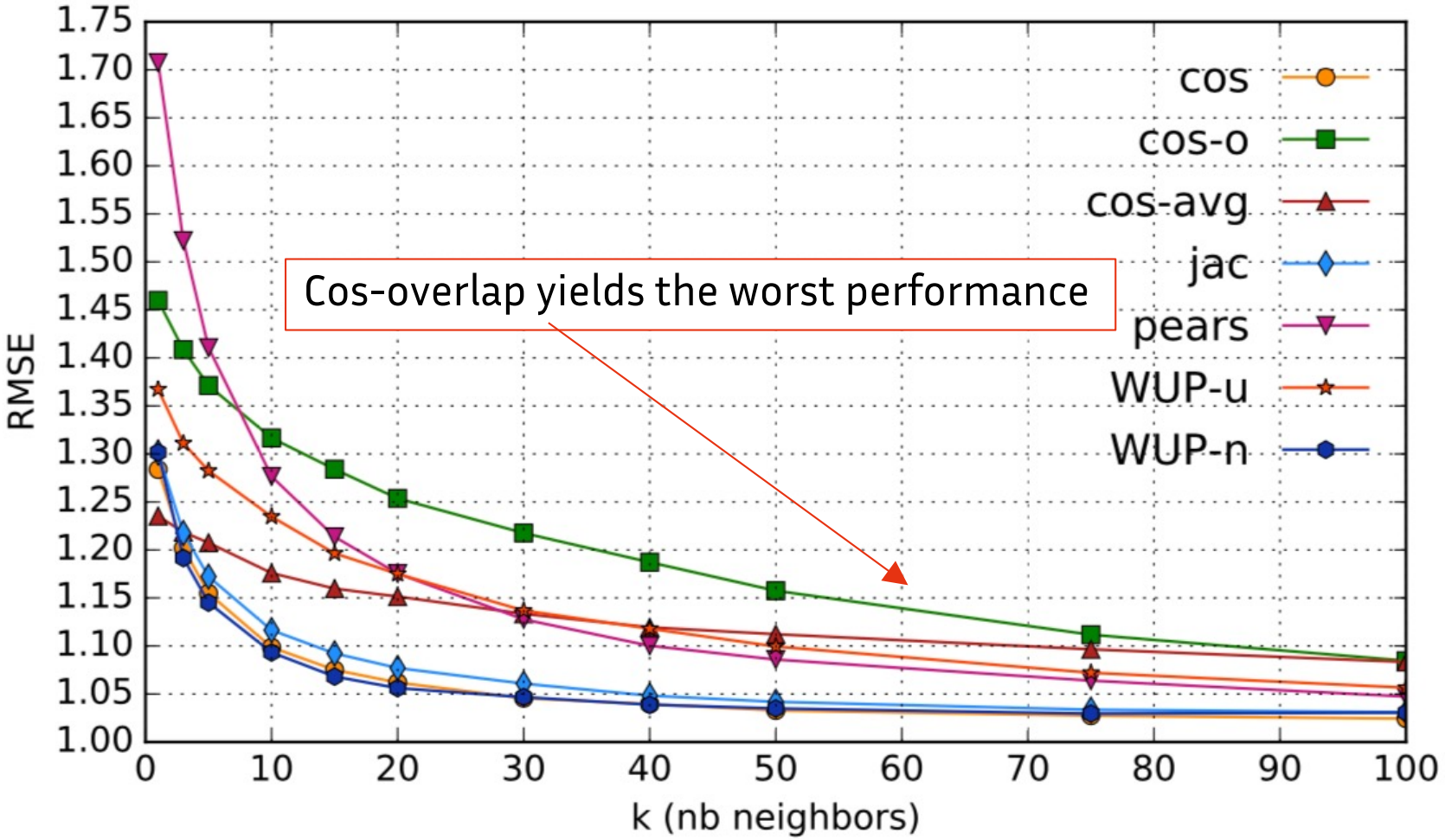
Results: Accuracy



Results: Expected Neighborhood



Results: RMSE



Summarizing Results

Cosine Overlap seems to be attack resilient, but offers poor recommendation performance.

- What makes it “resilient”?
- Can we keep resiliency while improving recommendation quality?

Understanding Sybil Resiliency

What makes Cos-Overlap so special?

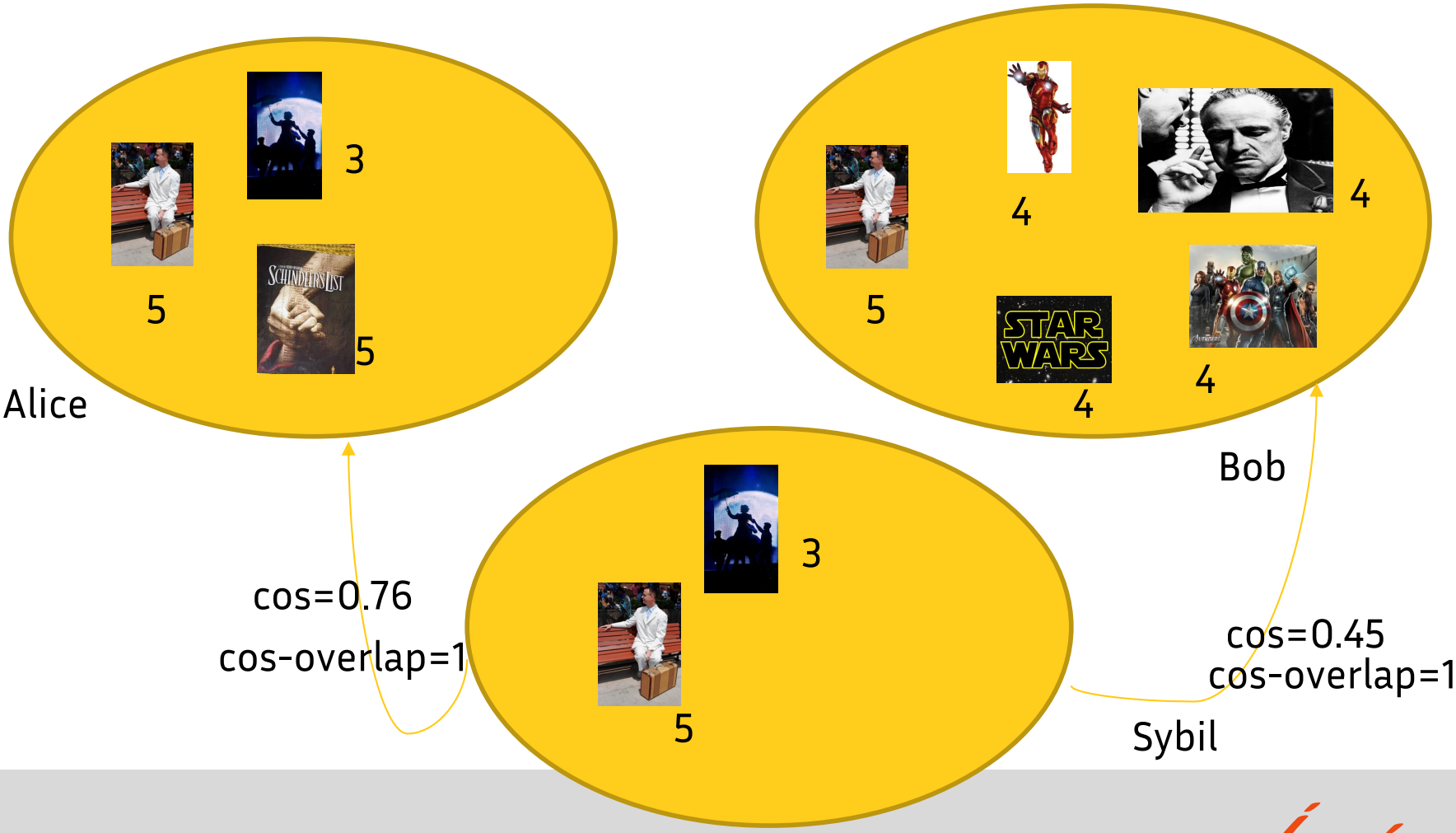
Cosine Similarity

$$Cos(u,n) = \frac{r_u \cdot r_n}{||r_u|| \cdot ||r_n||}$$

Cosine Overlap

$$Cos\text{-}overlap(u,n) = \frac{r_u \cdot r_n}{\sqrt{\sum_{i \in I_{u,n}} (r_{u,i})^2} \times \sqrt{\sum_{i \in I_{u,n}} (r_{n,i})^2}}$$

Understanding Sybil Resiliency



What Makes Cos-Overlap so Special?

Not normalizing by profile size

- Users that have same ratings on common items are perfectly similar
- Perfectly similar counterparts confuse Sybil neighborhoods.
- But they also hamper recommendation performance.

Intuition

Can we give perfectly similar counterparts only to Sybils?

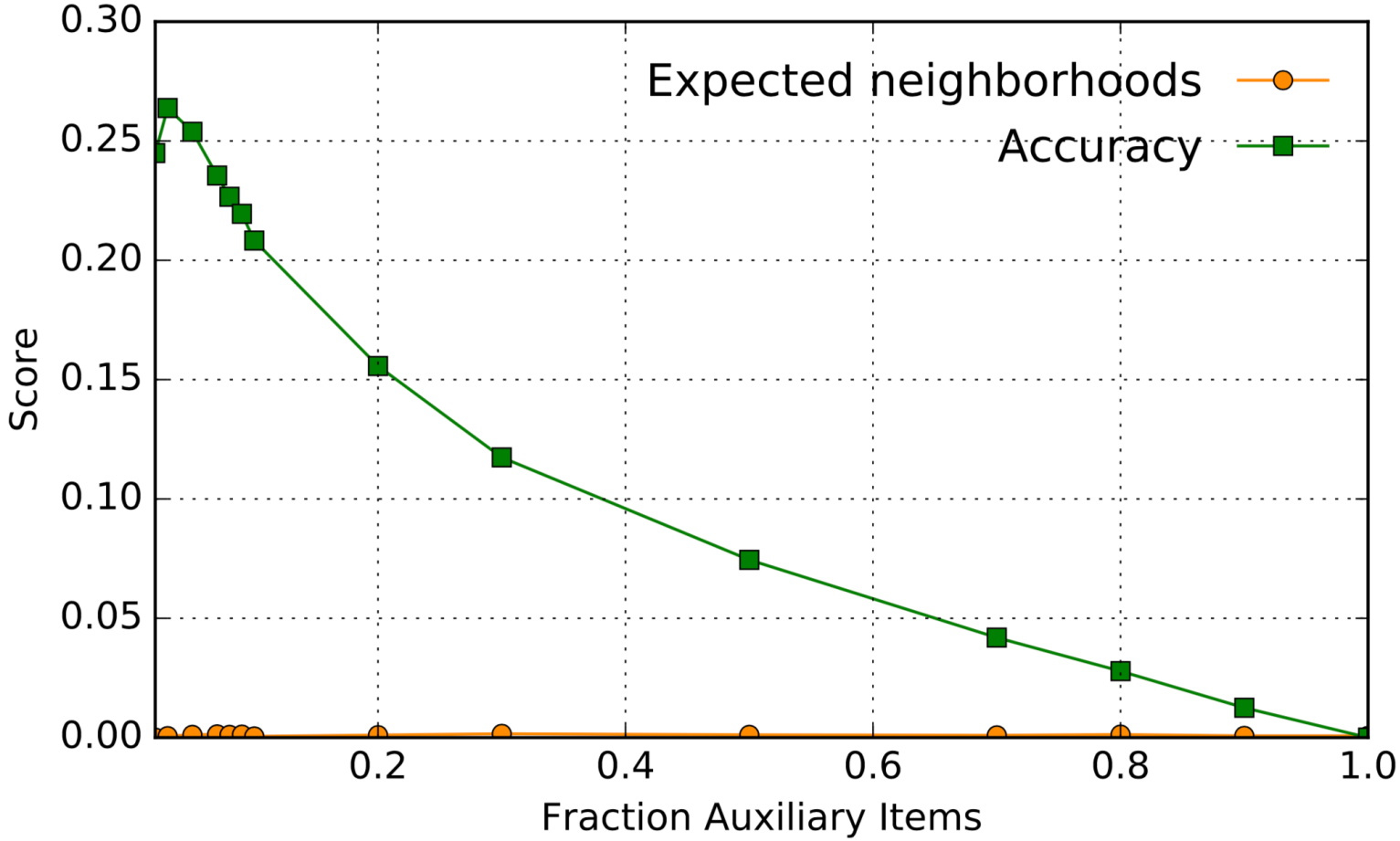
- Discriminate good from bad profiles for recommendation
- Prevent Sybils from identifying the target or other Sybils.

Two-Step Similarity Metric

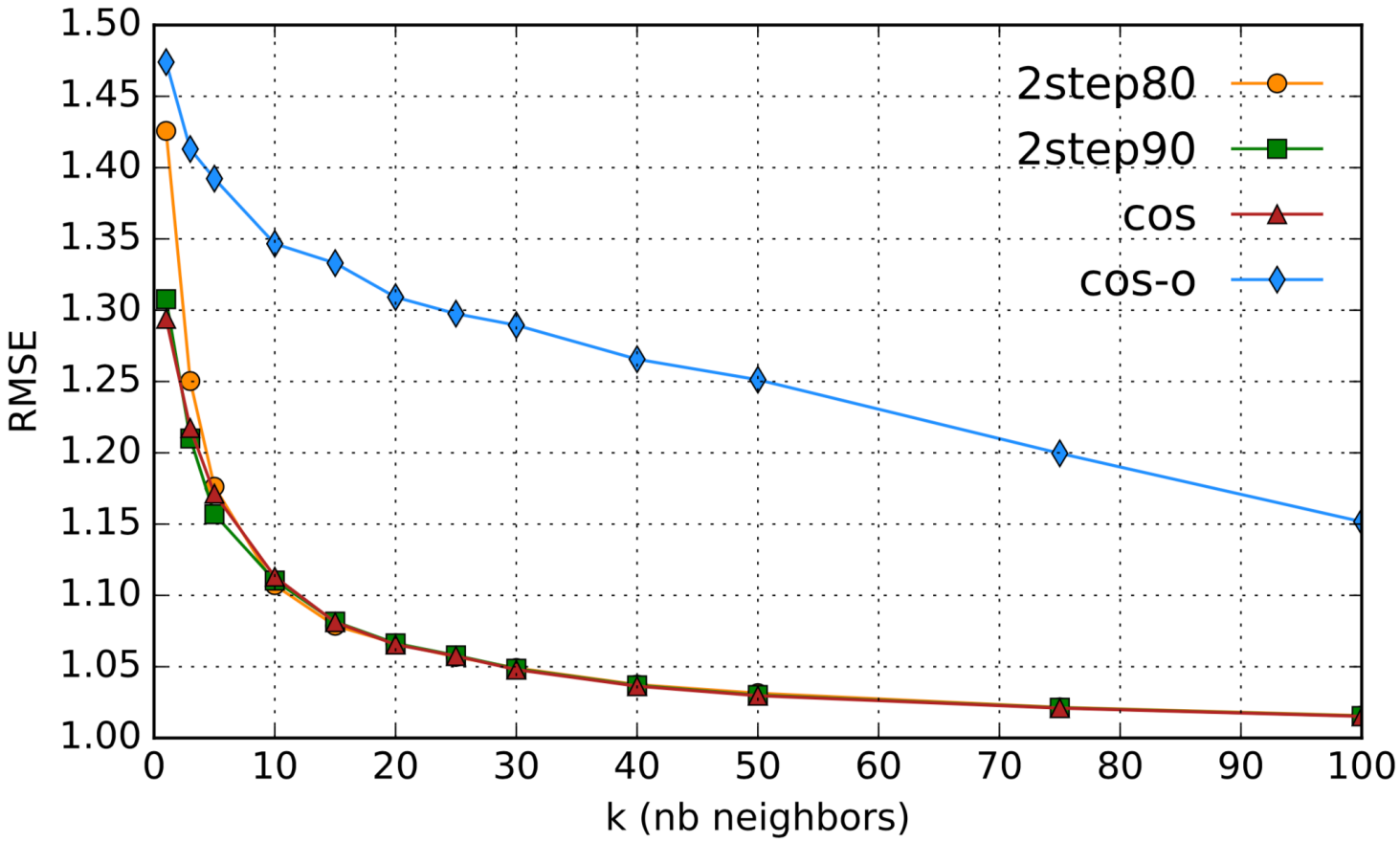
- STEP 1: Coalesce users that yield similar recommendation quality as neighbors
  - Cap similarity at a threshold *th*
- STEP 2: Privilege users that have the most items to recommend.
  - Similarity bonus for "new" items (items in *v* but not in *u*)

$$2\text{-}step(u,v) = \begin{cases} Sim(u,v) & \text{if } Sim(u,v) < th_u \\ th_u + f_{i,u}(|v-u|) & \text{if } Sim(u,v) \geq th_u \end{cases}$$

Two-Step Attack Resilience



Two Step Recommendation Quality



An Attack Targeting Two-Step

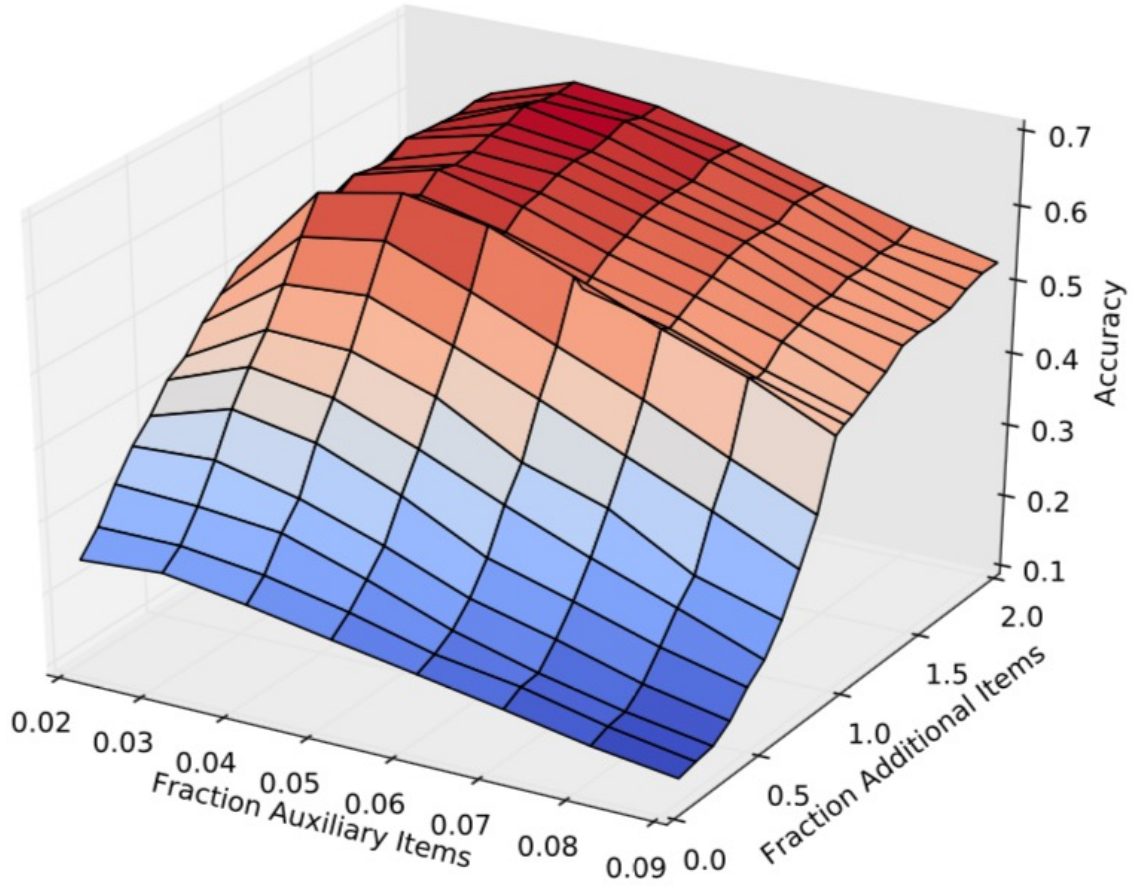
High similarity with 2-step implies

- a large enough set of common sitems,
- a large enough non-shared set.

New attack with

- set auxiliary items from the target's profile
- a set of additional items (fraction $q = \frac{\tilde{\mathcal{A}}\mathrm{dd}}{(1-p)|t|}$ |t|=size of target profile, p=fraction aux items)
- Additional items:
  - Unique/not unique
  - Fake or real

Accuracy of Specific Attack



Comparison with the State of the Art

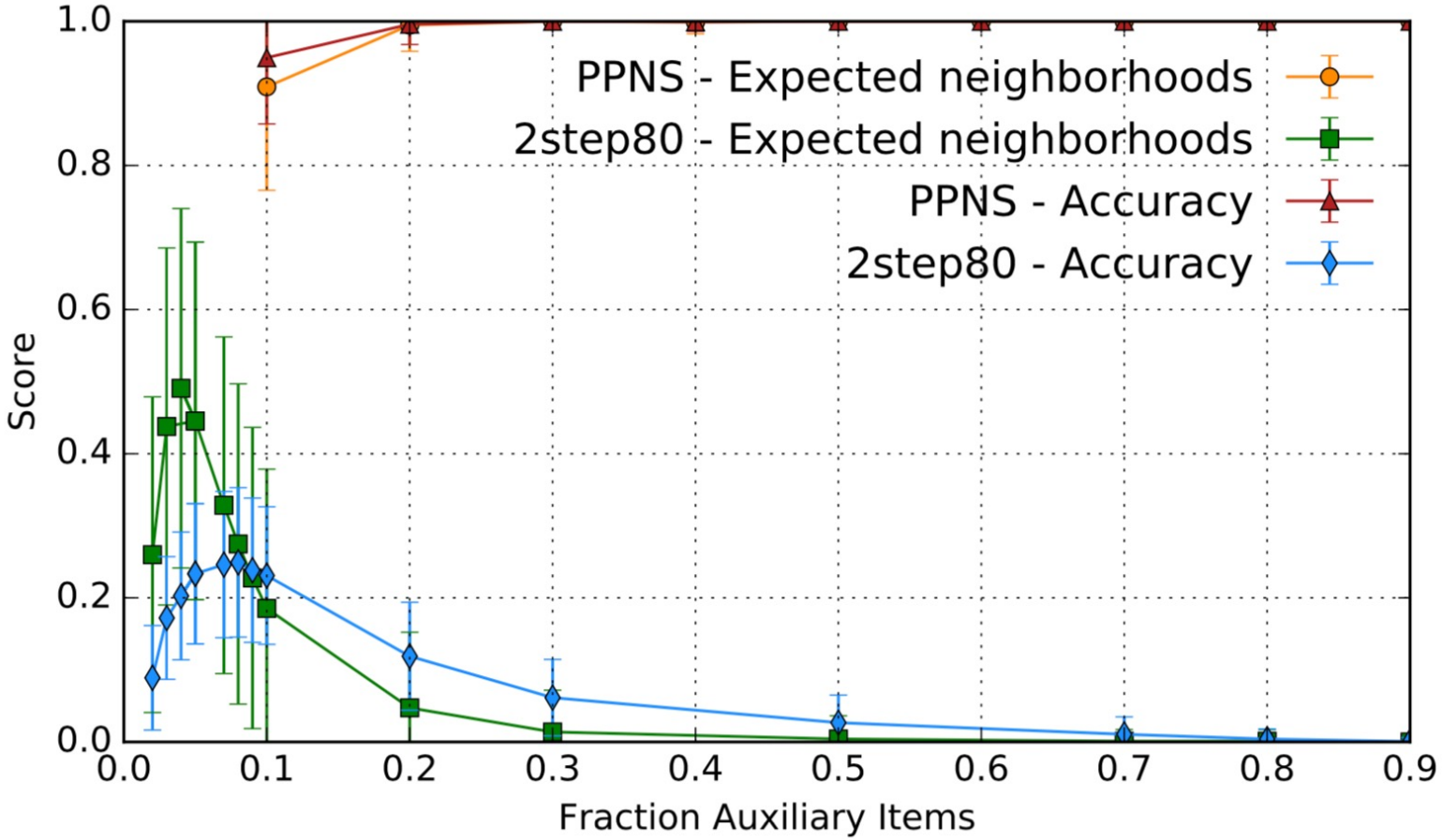
PPNS [LS14]:

- Partition potential neighbors by similarity classes
- Extract neighbors from top β partitions
- With at least one neighbor from β th one
- Works well with k Sybils,
- But what if we have β k?

[LS14] Z. Lu and H. Shen, "A security-assured accuracy-maximised privacy preserving collaborative filtering recommendation algorithm," in IDEAS, 2014

Comparison with State of the Art

Attack with Bk Sybils



Conclusions

- **Sybil Attack**
  - Effective on User Based Collaborative Filtering
- **Two-Step Similarity**
  - Effective Countermeasure with good recommendation quality
- **In the paper**
  - More experiments
  - Theoretical Analysis
- **Future Perspective**
  - Extend to other flavors of RS

Thank you!

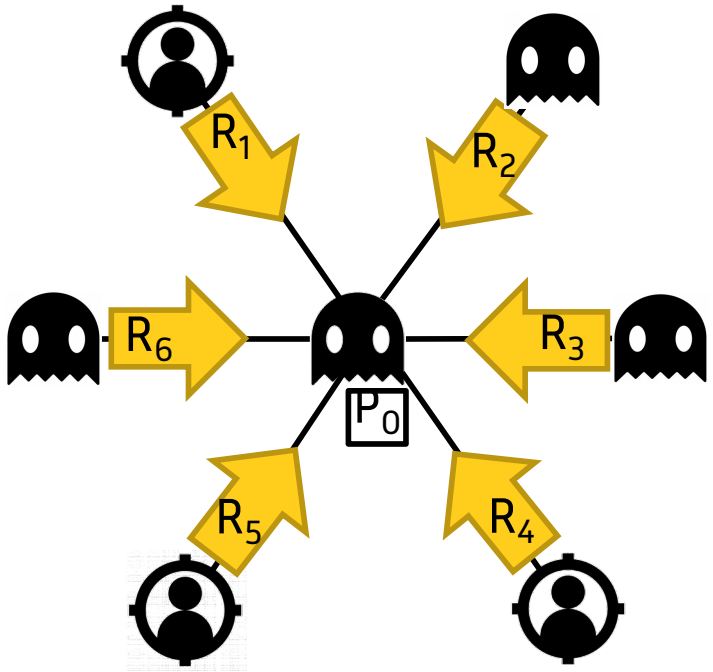
Follow us on www.inria.fr

inventors for the digital world

Yield

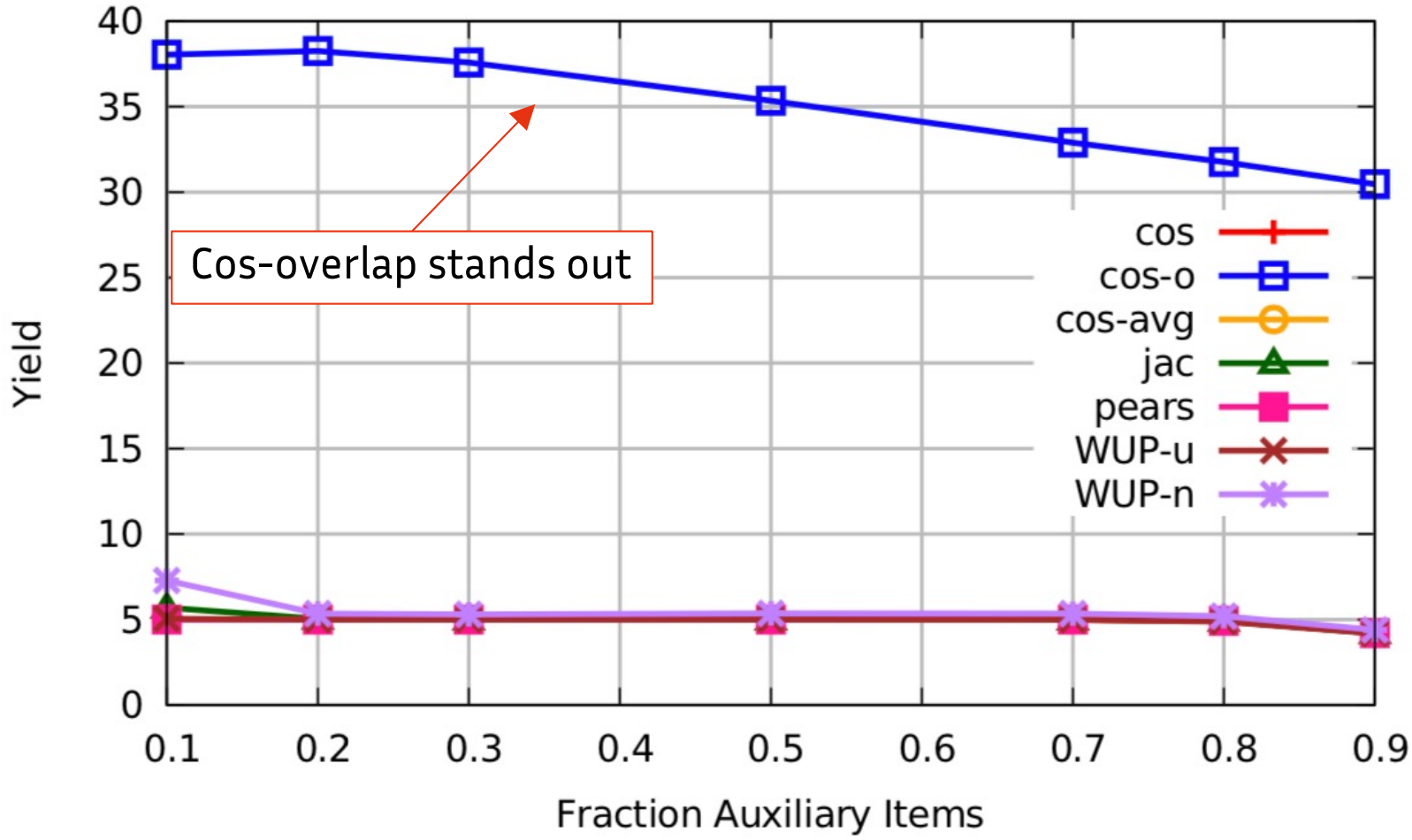
Each neighbor providers a user with a set of recommended items.

Yield measures the number of distinct and currently unrated recommended items by all of the Sybils together.

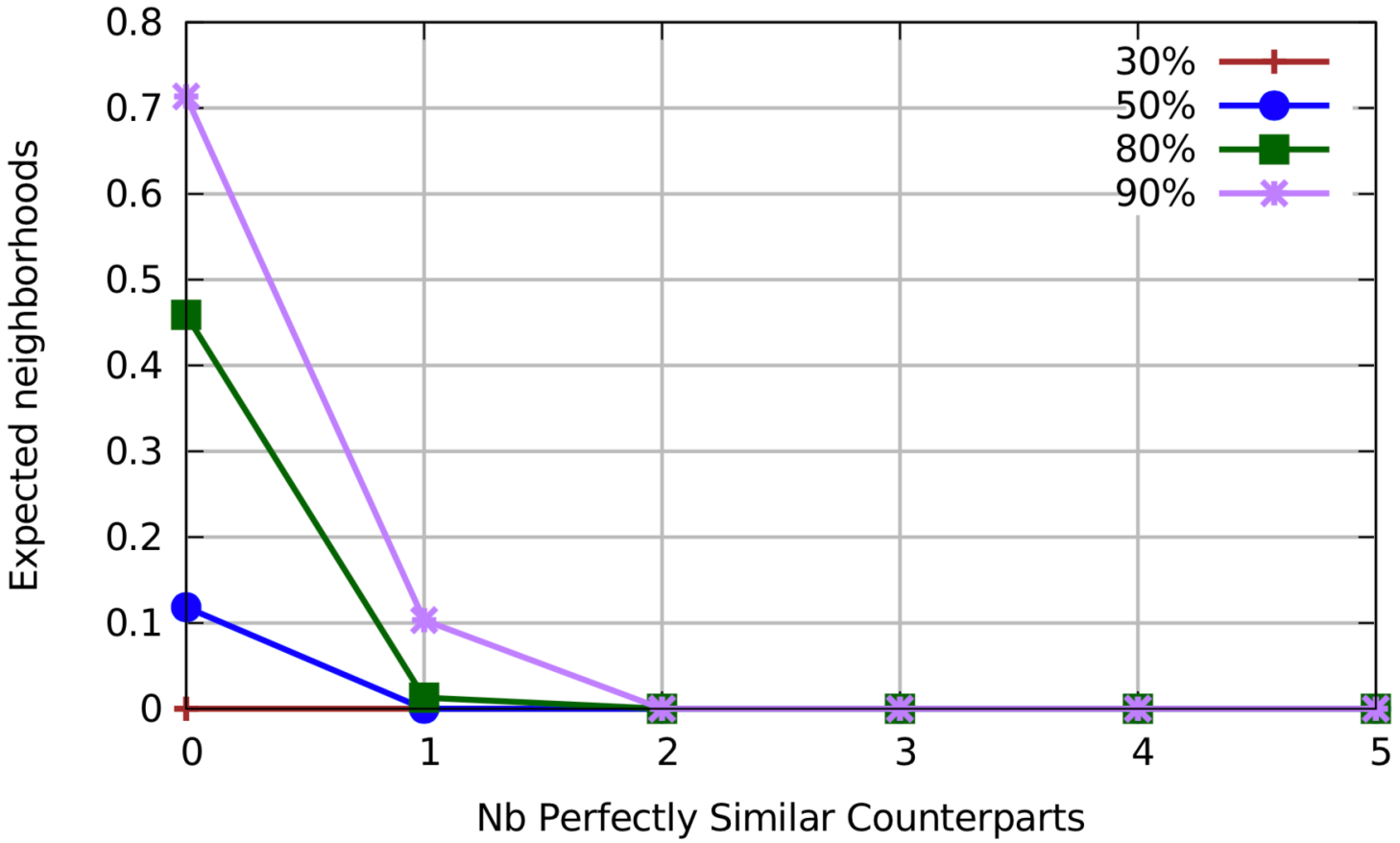


$$\text{Yield} = |\bigcup_{i=1}^{k} R_i \setminus P_0|$$

Results: Yield

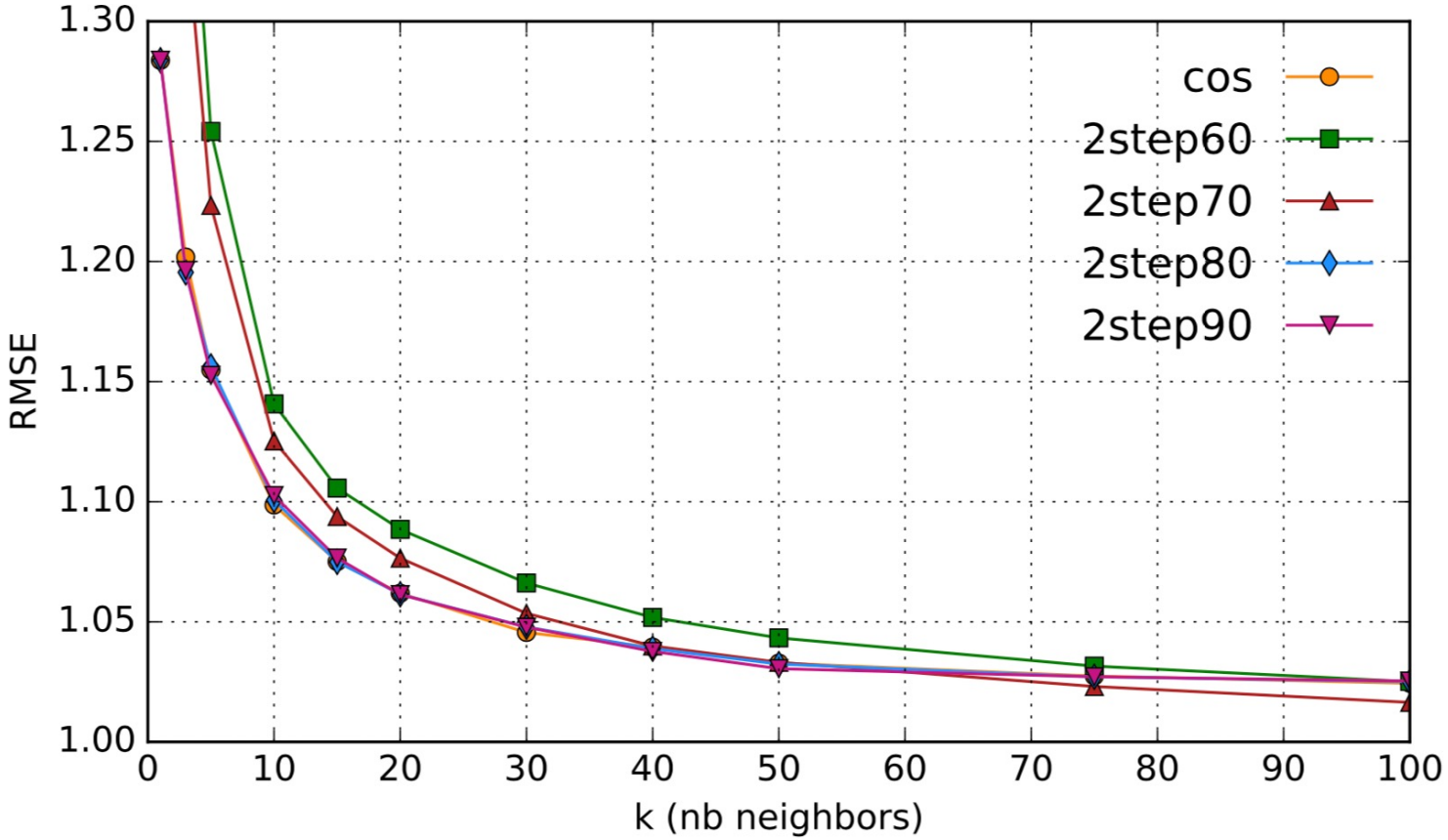


Attack vs Perfectly Similar Counterparts



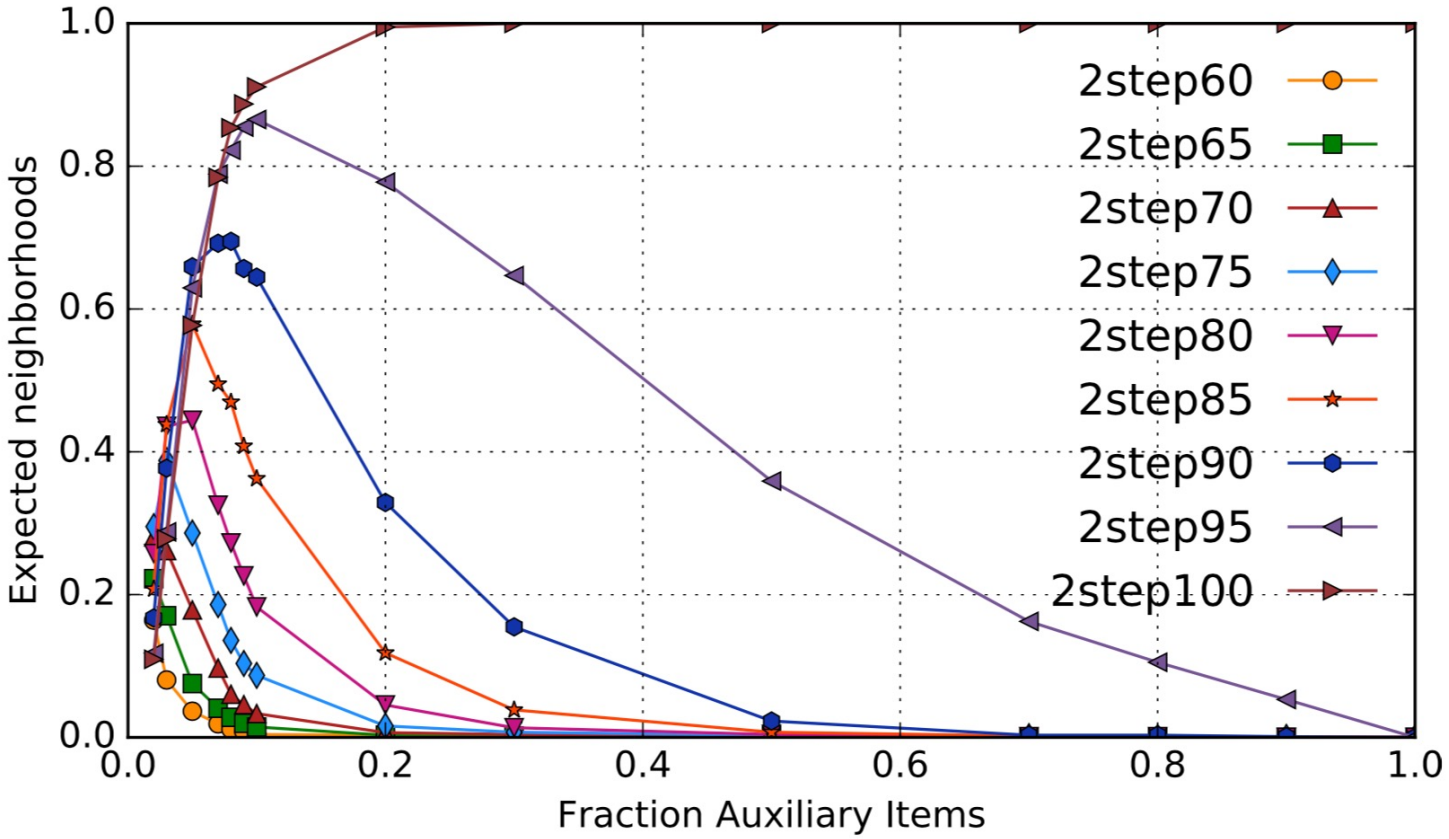
Impact of the Similarity Threshold

Recommendation Quality



Impact of the Similarity Threshold

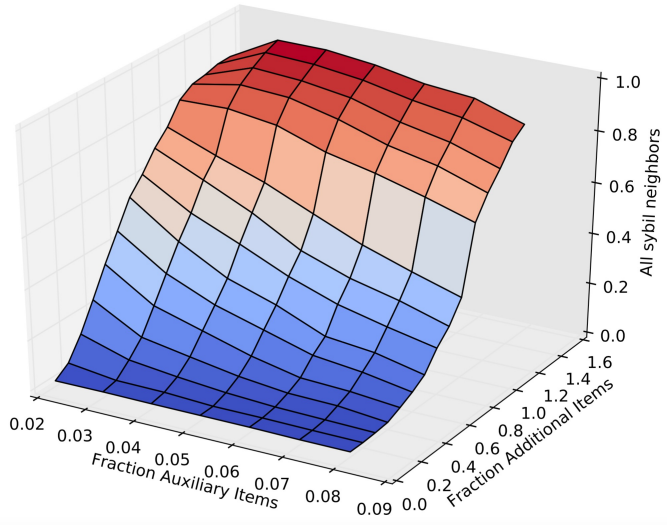
Attack Effectiveness

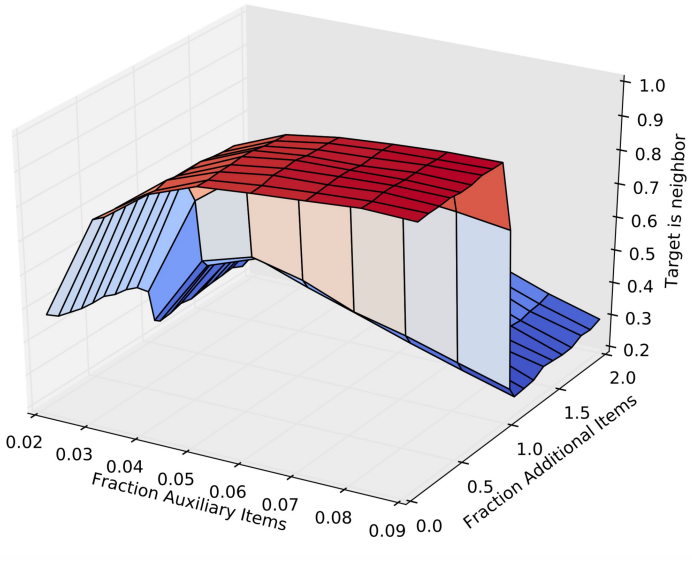


Expected Neighborhood

Consists of two components

- **All other sybils are neighbors**
- **Target is neighbor**

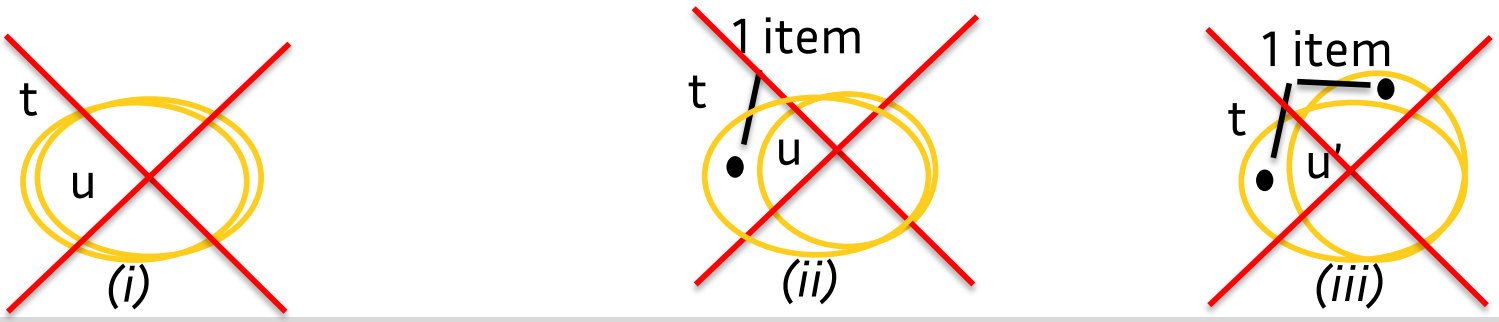




Theoretical Analysis

Let t be a target user with $|t| > 1$ and let $\mathcal{S}$ be the set of all Sybil users generated by the attacker.

Assumptions 1. *Let t be a user such that: (i) No other user has the same profile as t; (ii) there exists no user u with $u \subset t$ and $|t - u| = 1$; and (iii) there exists no user u' with $|u'| = |t|$ and $|u' \cap t| = |t| - 1$.*



Theoretical Analysis

Theorem 1. *Let s_{aux} be a Sybil user with a profile consisting of the auxiliary information aux. Given Assumptions 1, there exists aux that ensures: $\forall u \notin \mathcal{S}$, $cos(s,t) > cos(s,u)$*

Theorem 2. *Given Assumptions 1, let aux be such that: $aux \in t$, $|aux| = |t| - 1$, and $\forall u \subset t \implies u \subset aux$. Let s_{aux} be the associated Sybil user.*
Let $N_v = \#\{u \mid cos(u,v) \geq th_v\}$ be the number of users above the threshold for user v.
Let $C_v = \#\{u \mid u \subset v\}$ be the number of users with a profile which is a subset of v's profile. Then, if $C_t < N_{s_{aux}}$, we have: $\exists u \notin \mathcal{S}$, $2{-}step(s_{aux},t) \leq 2{-}step(s_{aux},u)$