Distributed Collaborative Filtering Protocol Based on Quasi-homomorphic Similarity

[†]Hiroaki Kikuchi, [†]**Yoshiki Aoki**,
[‡]Masayuki Terada, [‡]Kazuhiko Ishii, [‡]Kimihiko Sekino

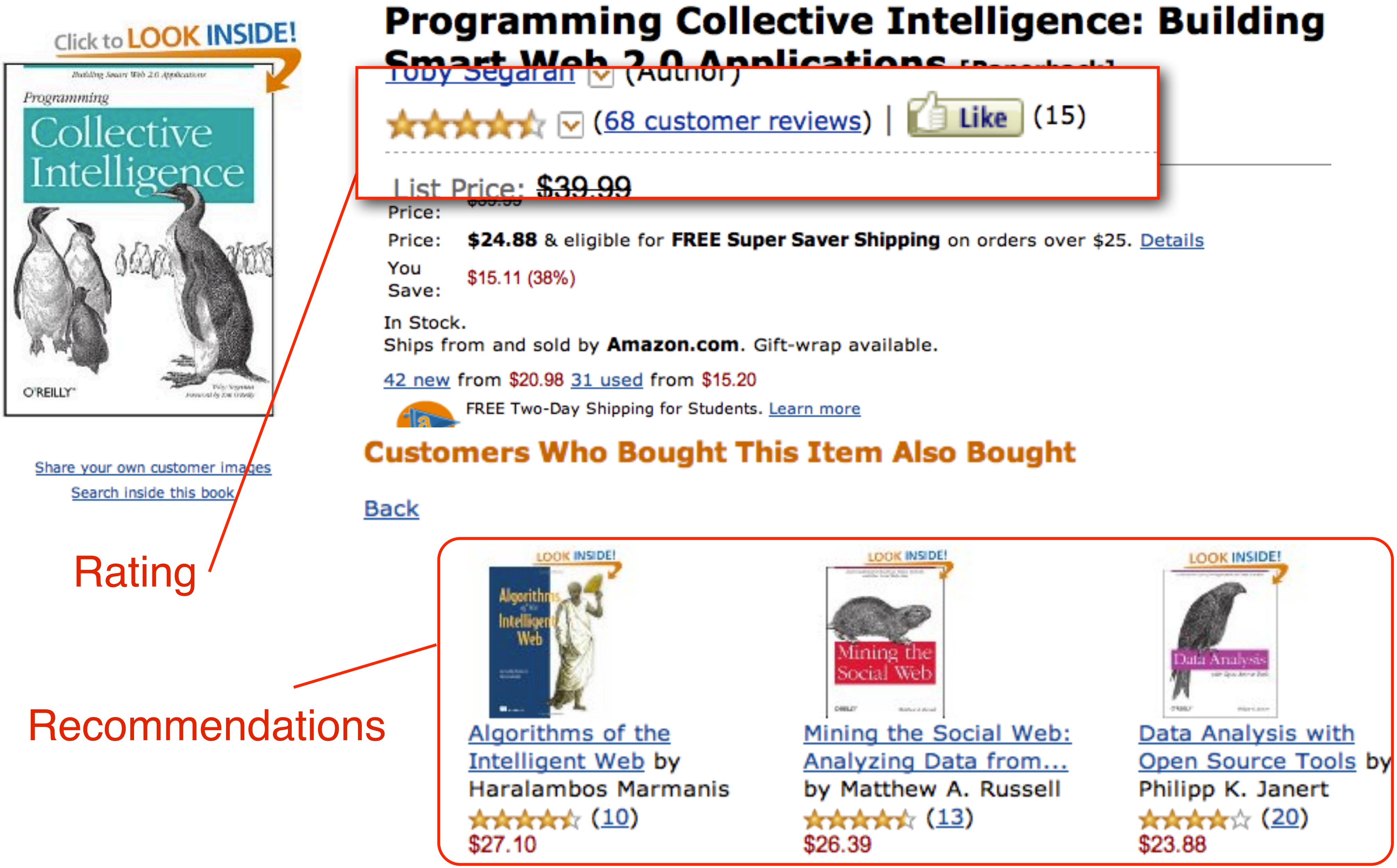
[†]Tokai University
[‡]NTT-DOCOMO

Presenter : Yoshiki Aoki .

Outline of my talk

1. Background
2. Our idea
3. Experiment
4. Conclusions

What is Recommendation?



Programming Collective Intelligence: Building Smart Web 2.0 Applications [Paperback]

Toby Segaran (Author)

(68 customer reviews) | Like (15)

List Price: $39.99

Price: $24.88 & eligible for FREE Super Saver Shipping on orders over $25. Details

You Save: $15.11 (38%)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

42 new from $20.98 31 used from $15.20

FREE Two-Day Shipping for Students. Learn more

Share your own customer images
Search inside this book

Customers Who Bought This Item Also Bought

Back

Algorithms of the Intelligent Web by Haralambos Marmanis (10) $27.10

Mining the Social Web: Analyzing Data from... by Matthew A. Russell (13) $26.39

Data Analysis with Open Source Tools by Philipp K. Janert (20) $23.88

Rating

Recommendations

Objective

amazon

A	i_1	i_2	Sim
u_1	4	3	1
u_2	1	4	0.69
u_3	1		0.13
u_4	4	3	-

ebay

B	i_3	i_4	i_5	Sim
u_1	5	1		0.12
u_2	3		4	0.29
u_3	4	5	2	0.59
u_4	*	2	2	-

We aim to get recommended items from the entire datasets.

Privacy Preserving Recommendation

Related works

Partition	Collaborative Filtering	other methods
Horizontal	Canny 02 (SVD) Kizawa 09 (Secure Intersection)	Sakuma 07(k-means) Clifton 04a(Association rule)
Vertical	**Our works**	Vaidya 03(Clustering) Kikuchi 10(NaïveBayes)

Two Approaches in Partitioning

Item ID

User ID

	i_1	i_2	i_3	i_4	i_5
u_1	3	2	5	1	
u_2	1	4	3		4
u_3	3		4	5	2
u_4	4	3		2	2

Horizontal

divided by user

3	2	5	1	
1	4	3		4

3		4	5	2
4	3		2	2

Vertical

divided by item

3	2
1	4
3	
4	3

5	1	
3		4
4	5	2
	2	2

What is Collaborative Filtering?

	i_1	i_2	i_3	i_4	i_5	Sim
u_1	3	2	5	1		0.9
u_2	1	4	3		4	0.2
u_3	3		4	5	2	0.4
u_4	4	3	*	2	2	-

Collaborative Filtering allows us to estimate any rating values based on the similarities between users.

$$\hat{r}^{AB}_{u,o} = \bar{r}_u + \frac{\sum_{v\in U-\{u\}} s_{u,v}(r_{v,i} - \bar{r}_v)}{\sum_{v\in U-\{u\}} s_{u,v}}$$

Outline of my talk

1. Background
2. Our idea
3. Experiment
4. Conclusions

Naive Collaborative Filtering (CF)

Joint comp. by A and B

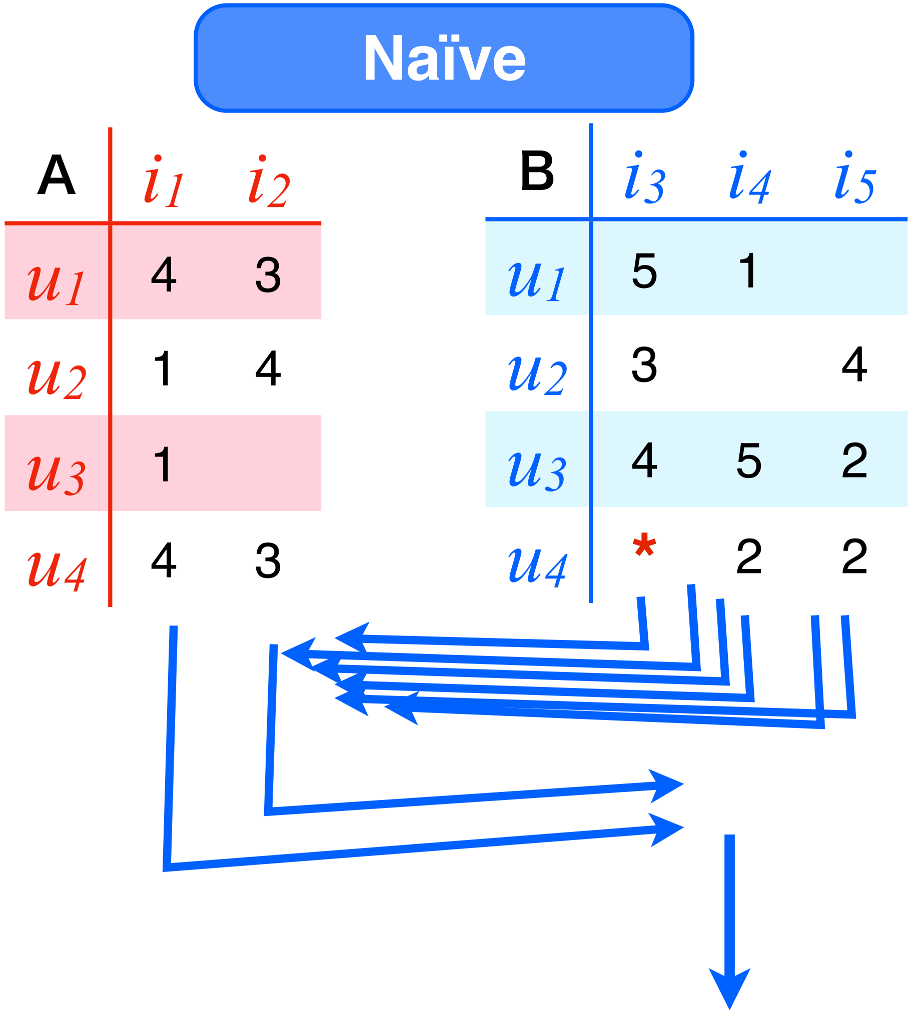
$$\frac{\begin{pmatrix} r_3 a_1 a_2 + \\ r_1 a_2 a_3 + \\ r_2 a_1 a_3 \end{pmatrix} + \begin{pmatrix} r_3 b_1 a_2 + r_2 b_1 a_3 + \\ r_3 b_2 a_1 + r_1 b_2 a_3 + \\ r_2 b_3 a_1 + r_1 b_3 a_2 \end{pmatrix} + \begin{pmatrix} r_3 b_1 b_2 + \\ r_1 b_2 b_3 + \\ r_2 b_1 b_3 \end{pmatrix}}{\begin{pmatrix} a_1 a_2 + \\ a_2 a_3 + \\ a_1 a_3 \end{pmatrix} + \begin{pmatrix} b_1 a_2 + b_1 a_3 + \\ b_2 a_1 + b_2 a_3 + \\ b_3 a_1 + b_3 a_2 \end{pmatrix} + \begin{pmatrix} b_1 b_2 + \\ b_2 b_3 + \\ b_1 b_3 \end{pmatrix}}$$

done by A

done by B

Problem expensive and complex $O(n^2)$

Our Idea

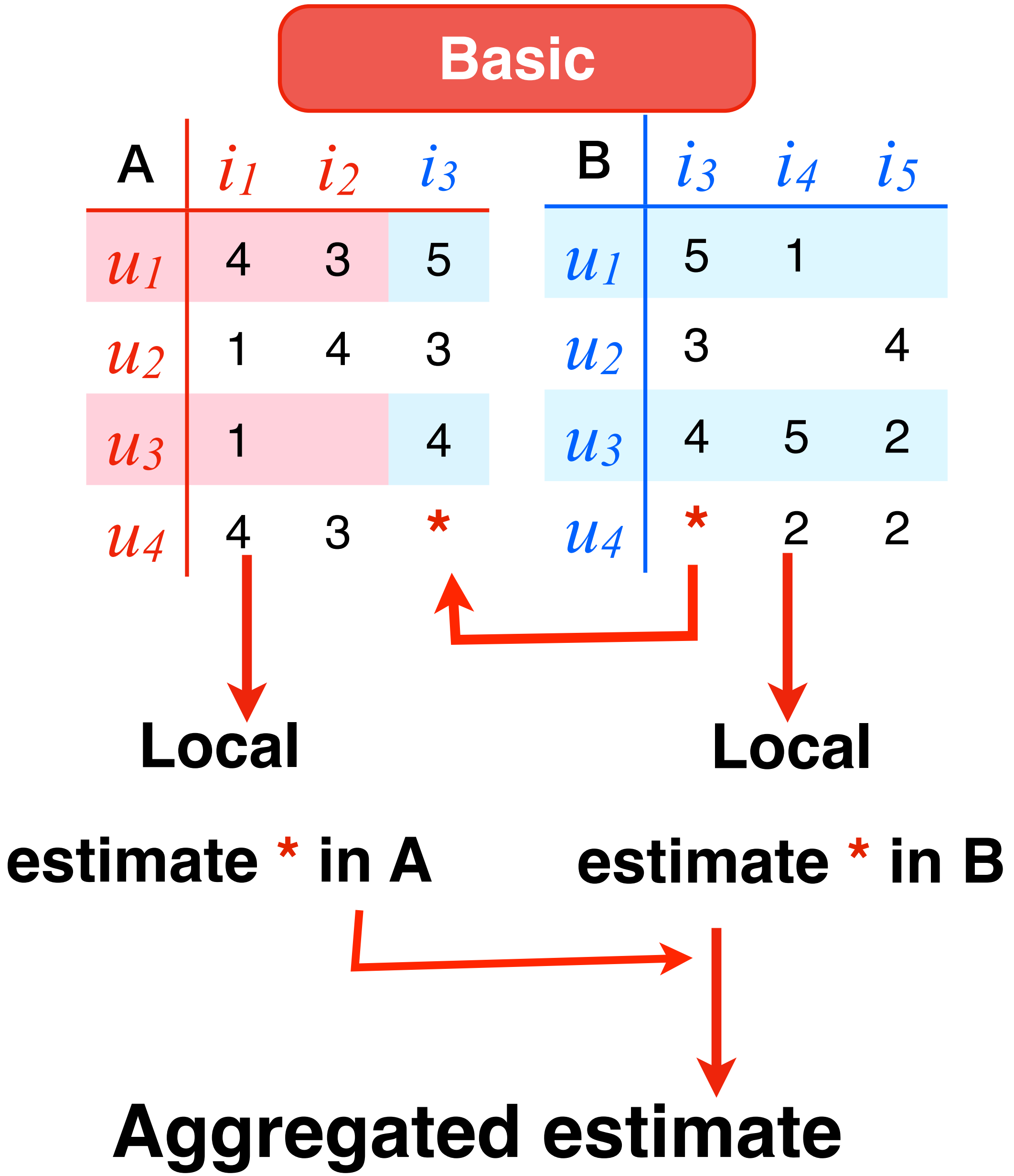


Naïve

A	i_1	i_2
u_1	4	3
u_2	1	4
u_3	1	
u_4	4	3

B	i_3	i_4	i_5
u_1	5	1	
u_2	3		4
u_3	4	5	2
u_4	*	2	2

Global estimate of *



Basic

A	i_1	i_2	i_3
u_1	4	3	5
u_2	1	4	3
u_3	1		4
u_4	4	3	*

B	i_3	i_4	i_5
u_1	5	1	
u_2	3		4
u_3	4	5	2
u_4	*	2	2

Local estimate * in A

Local estimate * in B

Aggregated estimate

Compare of complexity

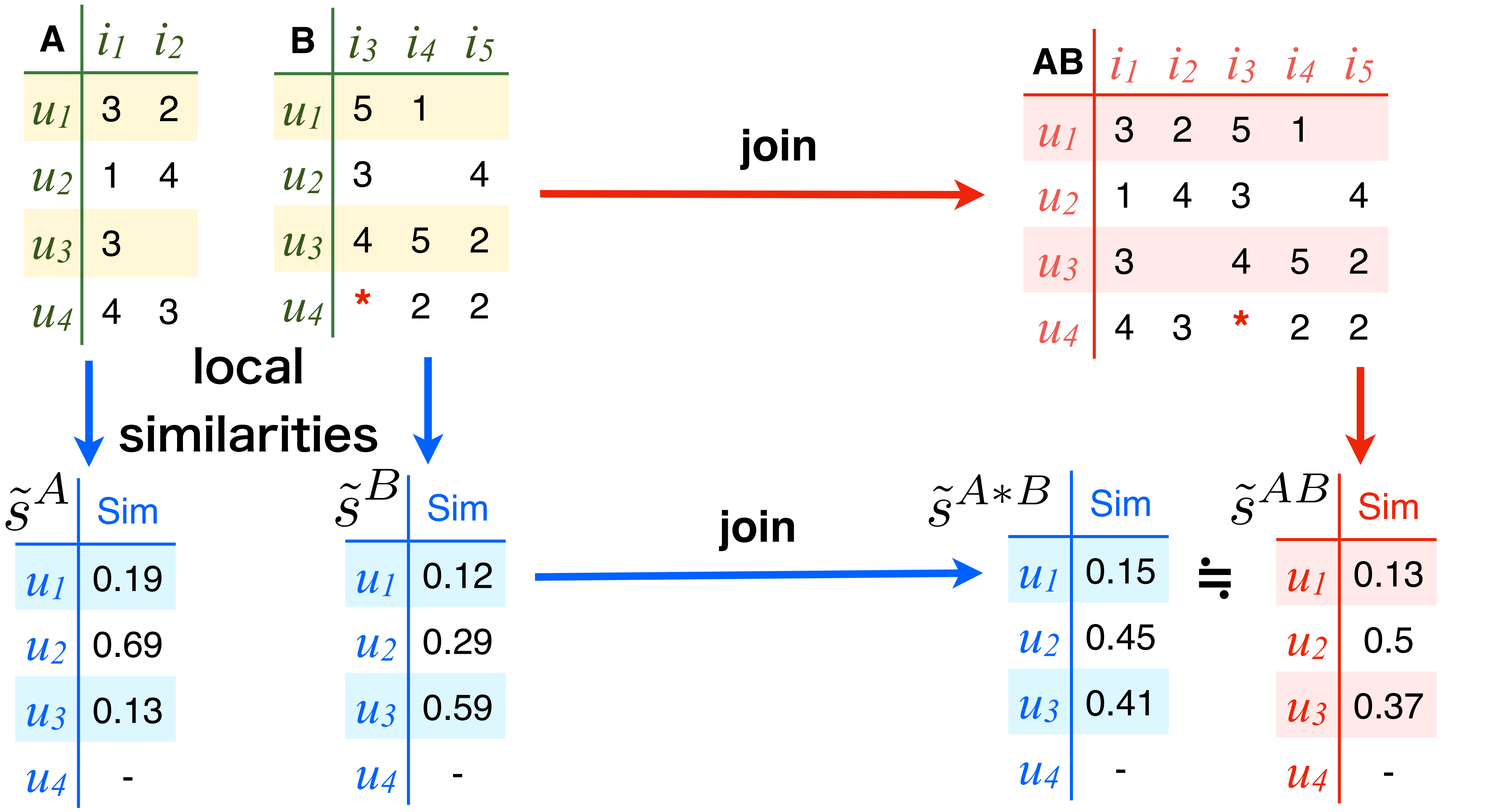
Naïve

$$\hat{r}^{AB}_{u,o} = \frac{\begin{pmatrix} r_3a_1a_2+ \\ r_1a_2a_3+ \\ r_2a_1a_3 \end{pmatrix} + \begin{pmatrix} r_3b_1a_2 + r_2b_1a_3+ \\ r_3b_2a_1 + r_1b_2a_3+ \\ r_2b_3a_1 + r_1b_3a_2 \end{pmatrix} + \begin{pmatrix} r_3b_1b_2+ \\ r_1b_2b_3+ \\ r_2b_1b_3 \end{pmatrix}}{\begin{pmatrix} a_1a_2+ \\ a_2a_3+ \\ a_1a_3 \end{pmatrix} + \begin{pmatrix} b_1a_2 + b_1a_3+ \\ b_2a_1 + b_2a_3+ \\ b_3a_1 + b_3a_2 \end{pmatrix} + \begin{pmatrix} b_1b_2+ \\ b_2b_3+ \\ b_1b_3 \end{pmatrix}}$$

Basic

$$\hat{r}^{A*B}_{u,o} = \frac{\begin{pmatrix} r_1a_2a_3+ \\ r_2a_1a_3+ \\ r_3a_1a_2 \end{pmatrix}}{\begin{pmatrix} a_2a_3+ \\ a_1a_3+ \\ a_1a_2 \end{pmatrix}} w_A + \frac{\begin{pmatrix} r_3b_1b_2+ \\ r_1b_2b_3+ \\ r_2b_1b_3 \end{pmatrix}}{\begin{pmatrix} b_1b_2+ \\ b_2b_3+ \\ b_1b_3 \end{pmatrix}} w_B$$

Quasi-homomorphic similarity



A	i_1	i_2
u_1	3	2
u_2	1	4
u_3	3	
u_4	4	3

B	i_3	i_4	i_5
u_1	5	1	
u_2	3		4
u_3	4	5	2
u_4	*	2	2

join

AB	i_1	i_2	i_3	i_4	i_5
u_1	3	2	5	1	
u_2	1	4	3		4
u_3	3		4	5	2
u_4	4	3	*	2	2

local similarities

$\tilde{s}^A$	Sim
u_1	0.19
u_2	0.69
u_3	0.13
u_4	-

$\tilde{s}^B$	Sim
u_1	0.12
u_2	0.29
u_3	0.59
u_4	-

join

$\tilde{s}^{A*B}$	Sim
u_1	0.15
u_2	0.45
u_3	0.41
u_4	-

$\fallingdotseq$

$\tilde{s}^{AB}$	Sim
u_1	0.13
u_2	0.5
u_3	0.37
u_4	-

Quasi homomorphic Similarity

$$|\tilde{s}^{A*B} - \tilde{s}^{AB}| < \epsilon$$

Proposed algorithms

- Algorithm 1: Basic scheme
- Algorithm 2: Pre-computation
- Algorithm 3: k-Nearest neighbor

Algorithm 1. Basic Scheme

Step2

$$E[5]^{0.4} \cdot E[3]^{0.1} \cdot E[4]^{0.5} = E[4.3]$$

A	i_1	i_2	i_3	$\tilde{s}$
u_1	3	2	[5]	0.4
u_2	1	4	[3]	0.1
u_3	3		[4]	0.5
u_4	4	3	*	-
			Sum	1

Step1

$(4, E[5], E[3], E[4])$

Step3

$(E[4.3])$

B	i_3	i_4	i_5	$\tilde{s}$
u_1	5	1		0.6
u_2	3		4	0.3
u_3	4	5	2	0.1
u_4	*	2	2	-
			Sum	1

Step4

$$\hat{r}^{A*B}_{4,3} = D[E[4.3]] \; \frac{2}{5} \; + \; 4.4 \; \frac{3}{5} = 4.36$$

Problem: Performance

	computation costs	processing time [ms]
$E[m]$	**n-1**	**168**
$E[m_1] \cdot E[m_2]$	n-1	0.102
$E[m_1]^{m_2}$	n-1	0.093
overall approximation time(n=943)		158 [s]

The exponents are limited within relatively small numbers in CF and hence the processing time is extremely smaller than that of an ordinary modular exponentiations with exponent chosen from full domain.

Algorithm 2. Pre-computation

- **Prepare ciphertexts** of rating in advance.

$$R' = \{E[1], E[2], E[3], E[4], E[5]\}$$

- Prepare q ciphertexts of zero($E[0]_i \neq E[0]_j$.)

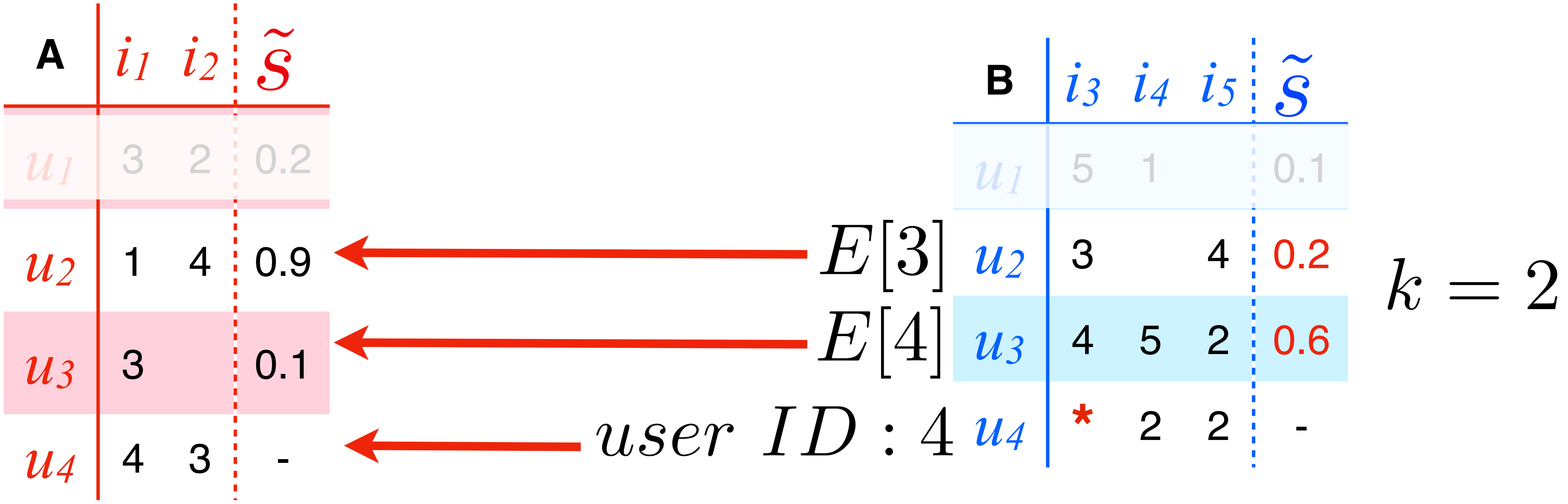
$$Z = \{E[0]_1, \ldots, E[0]_q\}$$

- Generate a new ciphertext from Z

$$E[r]' = E[r] \cdot E[0]$$

Algorithm 3. k-Nearest Neighbor

- We restrict users within the k nearest users.
  - Example) $k = 2$



A	i_1	i_2	$\tilde{s}$
u_1	3	2	0.2
u_2	1	4	0.9
u_3	3		0.1
u_4	4	3	-

$E[3]$

$E[4]$

$user\ ID : 4$

B	i_3	i_4	i_5	$\tilde{s}$
u_1	5	1		0.1
u_2	3		4	0.2
u_3	4	5	2	0.6
u_4	*	2	2	-

$k = 2$

B chooses top 2 users in the order of similarity.

Outline of my talk

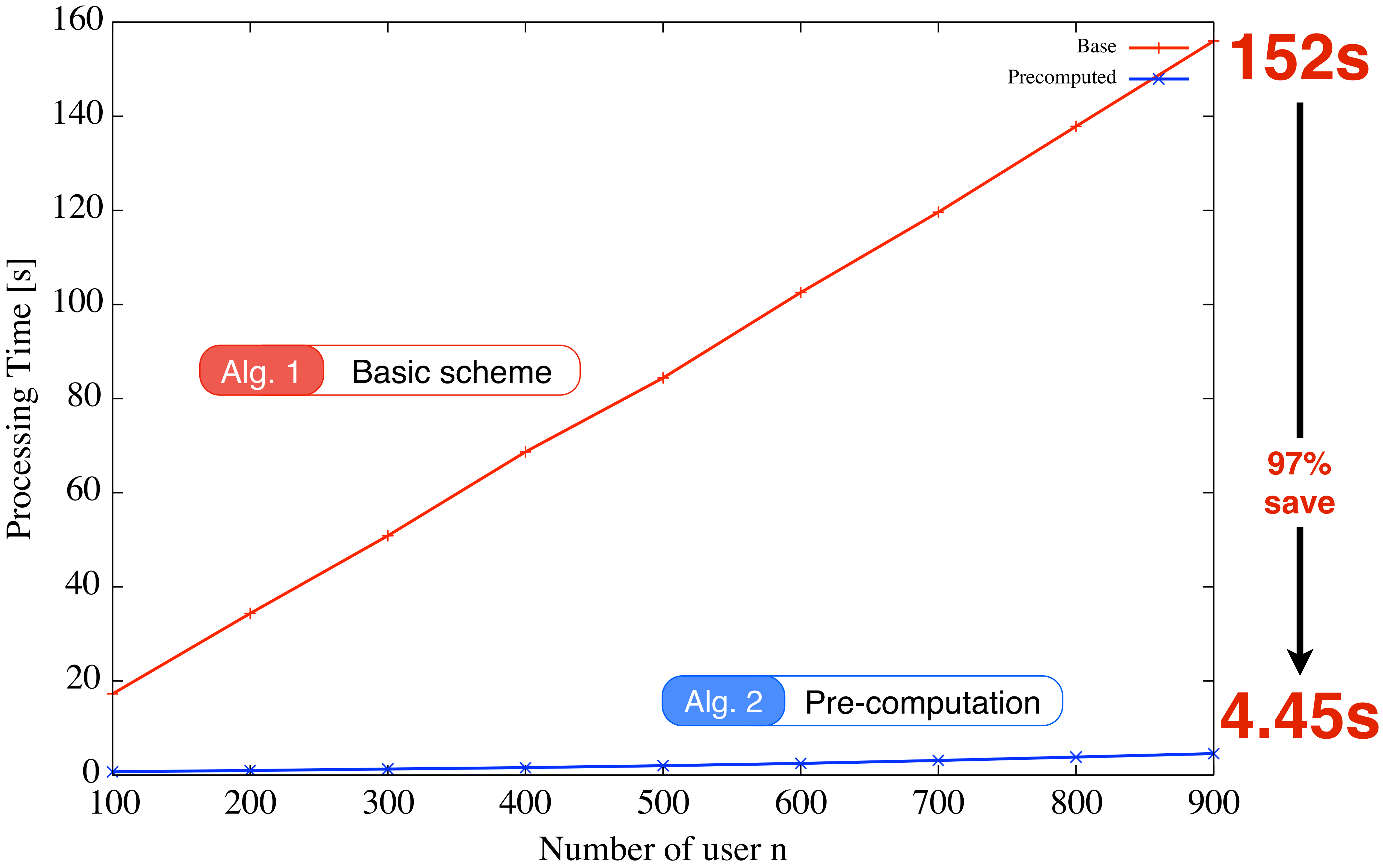
1. Background
2. Our idea
3. Experiment
4. Conclusions

Experiment

1. Computation time.
2. Comparison of similarities.
3. Accuracy of prediction.

- Intel Core 2 Duo 2.26GHz, 4GB, Java version 1.6.
- Dataset: MovieLens Data set
  - ratings: **100,000**
  - users: **943**
  - items: **1,628**

1. Processing Time



Alg. 1 Basic scheme

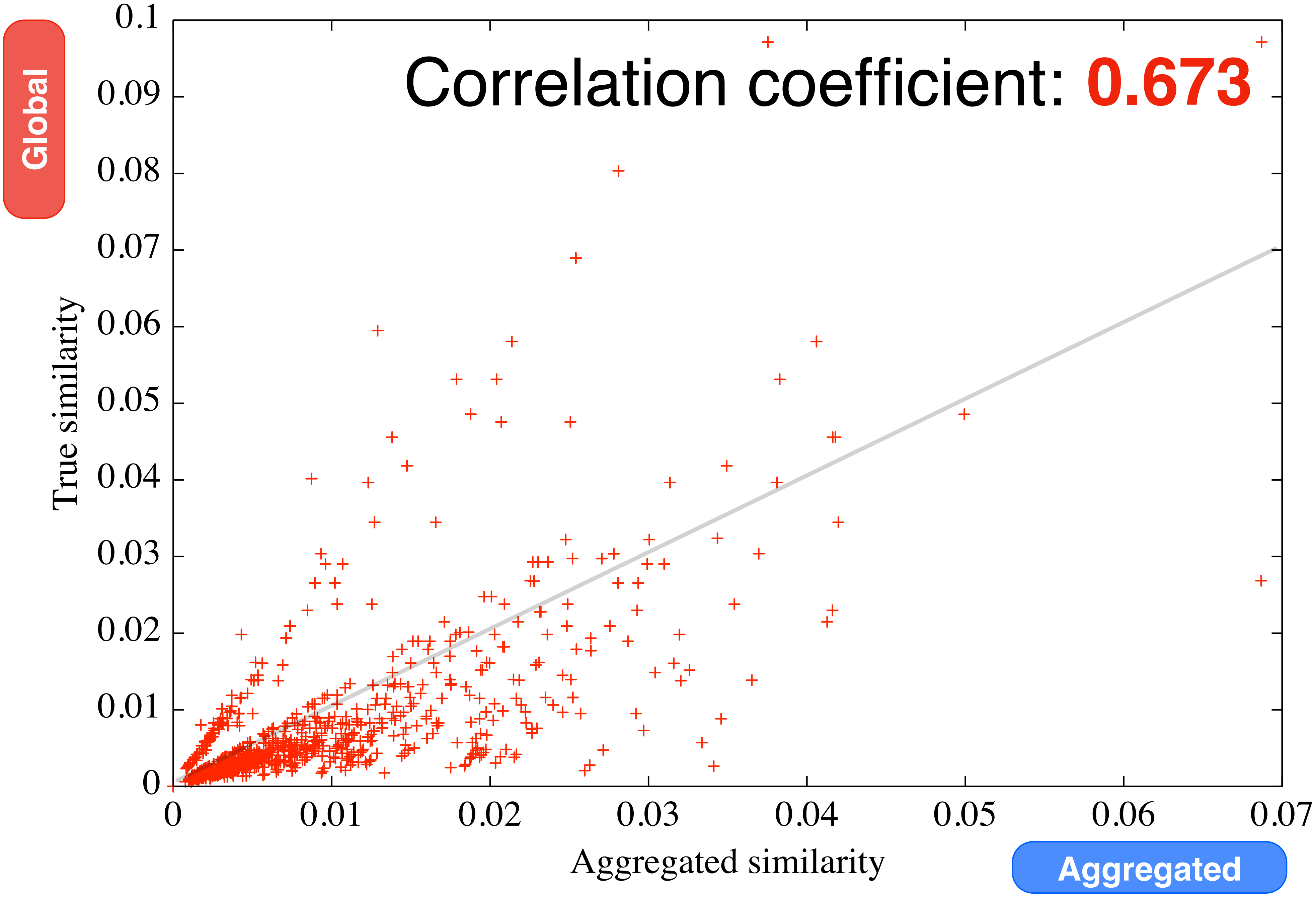
Alg. 2 Pre-computation

152s

97% save

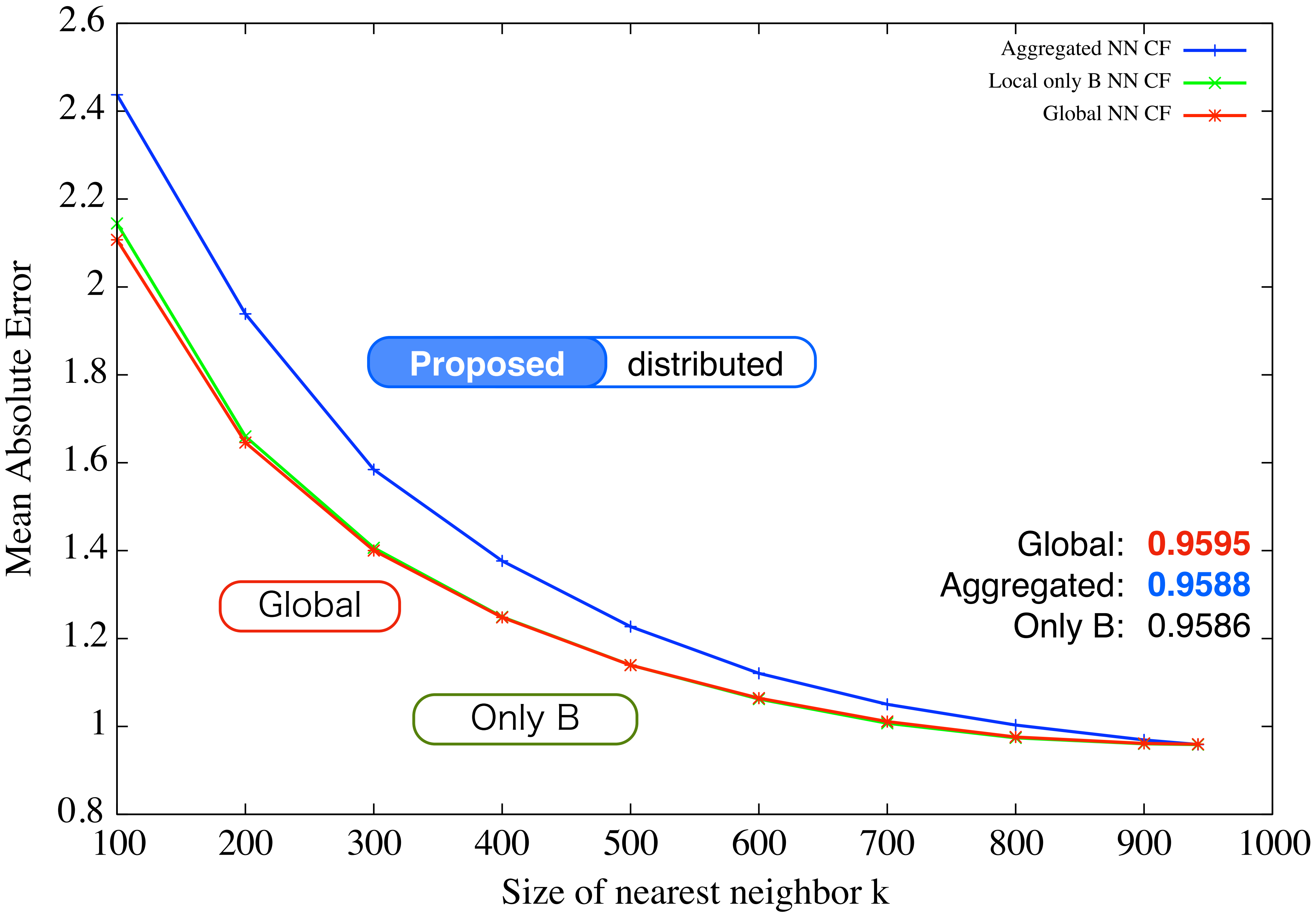
4.45s

2. Comparison of similarities



Correlation coefficient: **0.673**

3. Accuracy (MAE)



Global: 0.9595
Aggregated: 0.9588
Only B: 0.9586

Conclusions

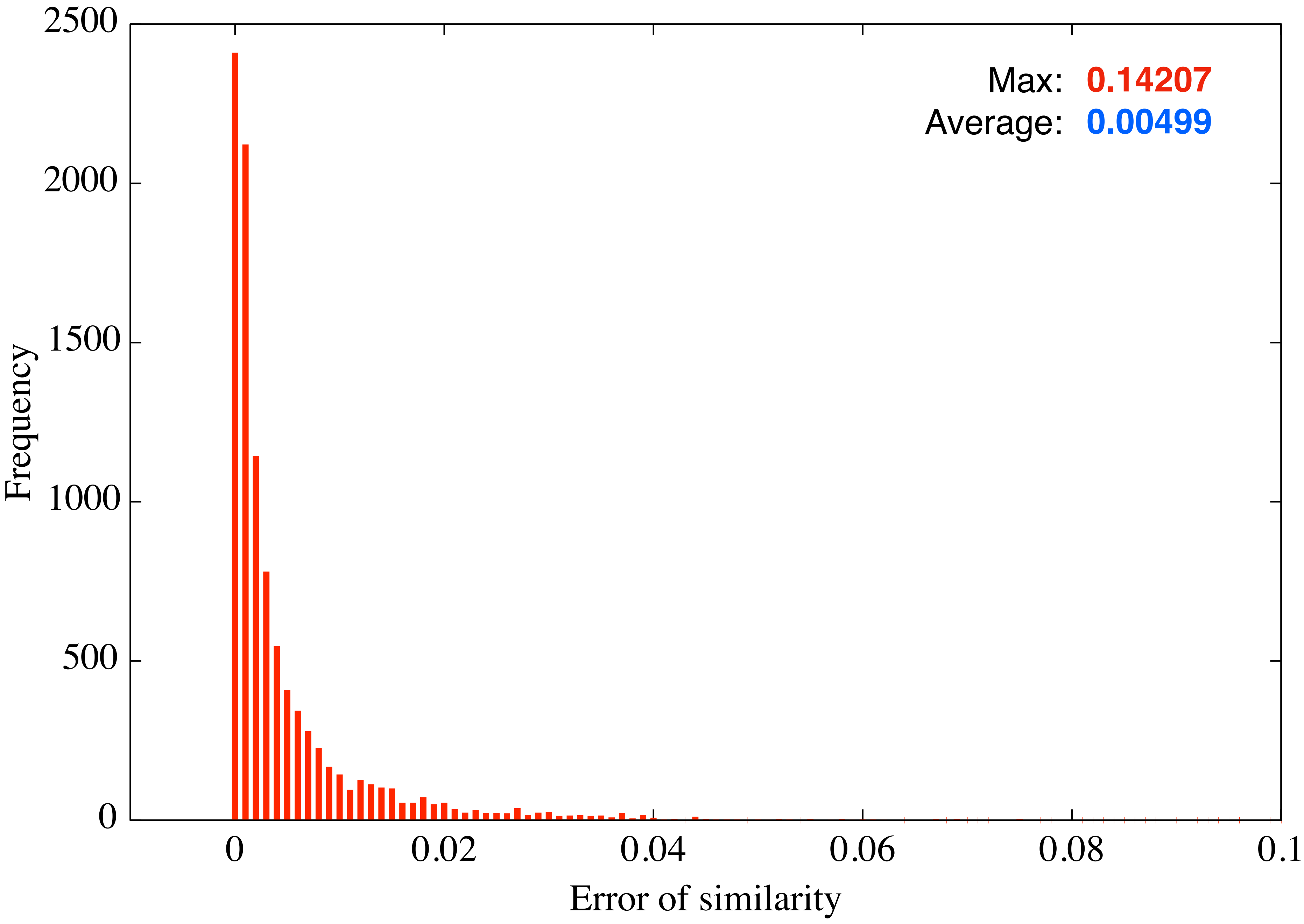
- We have proposed a new **private** CF protocol from **vertically** partitioned datasets using the **quasi-homomorphic similarity**.

- This table gives the summary of our proposed schemes. All of three schemes are excellent performance in terms of communication and computation costs, though prediction accuracies are fair. We conclude that the best scheme is the combination of the Pre-computation and the nearest neighbor scheme.

	Computation cost	Communication cost	Accuracy
Naive	High	High	Good
Alg. 1: Basic scheme	Low	Low	Fair
Alg. 2: Pre-computation	Excellent	Low	Fair
Alg. 3: -Nearest neighbor	Excellent	Excellent	Fair

- Our future studies include the optimal similarity and the treatment of missing values.

Histogram of difference between the two similarities.



Similarity

- 1/(Euclidean Distance + 1)

$$s_{u,v} = \frac{1}{1+\sum_{i\in I_u\cap I_v}(r_{u,i}-r_{v,i})^2}$$

- Example

	i_1	i_2	i_3	i_4	i_5
u_1	3	2	5	1	
u_2	1	4	3		4

$$s_{1,2} = \frac{1}{1+(3-1)^2+(2-4)^2+(5-3)^2} = 0.0769$$

Extended CF equation

- dropped some eletmets

$$\hat{r}_{u,i} = \cancel{\bar{r}_u} + \frac{\sum_{v \in U - \{u\}} s_{u,v}(r_{v,i} - \cancel{\bar{r}_v})}{\sum_{v \in U - \{u\}} s_{u,v}}$$

$$\hat{r}_{u,i} = \frac{\sum_{v \in U - \{u\}} s_{u,v} r_{v,i}}{\sum_{v \in U - \{u\}} s_{u,v}}$$

Paillier encryption

- equation of encrypt

$$E[m] = g^m r^n \pmod{n^2}$$

- additive homomorphic property

$$E[m_1] \cdot E[m_2] = E[m_1 + m_2]$$

$$E[m_1]^{m_2} = E[m_1 \cdot m_2]$$

- example

$$E[5] \cdot E[3] = E[5 + 3] = E[8]$$

$$E[5]^3 = E[5 * 3] = E[15]$$

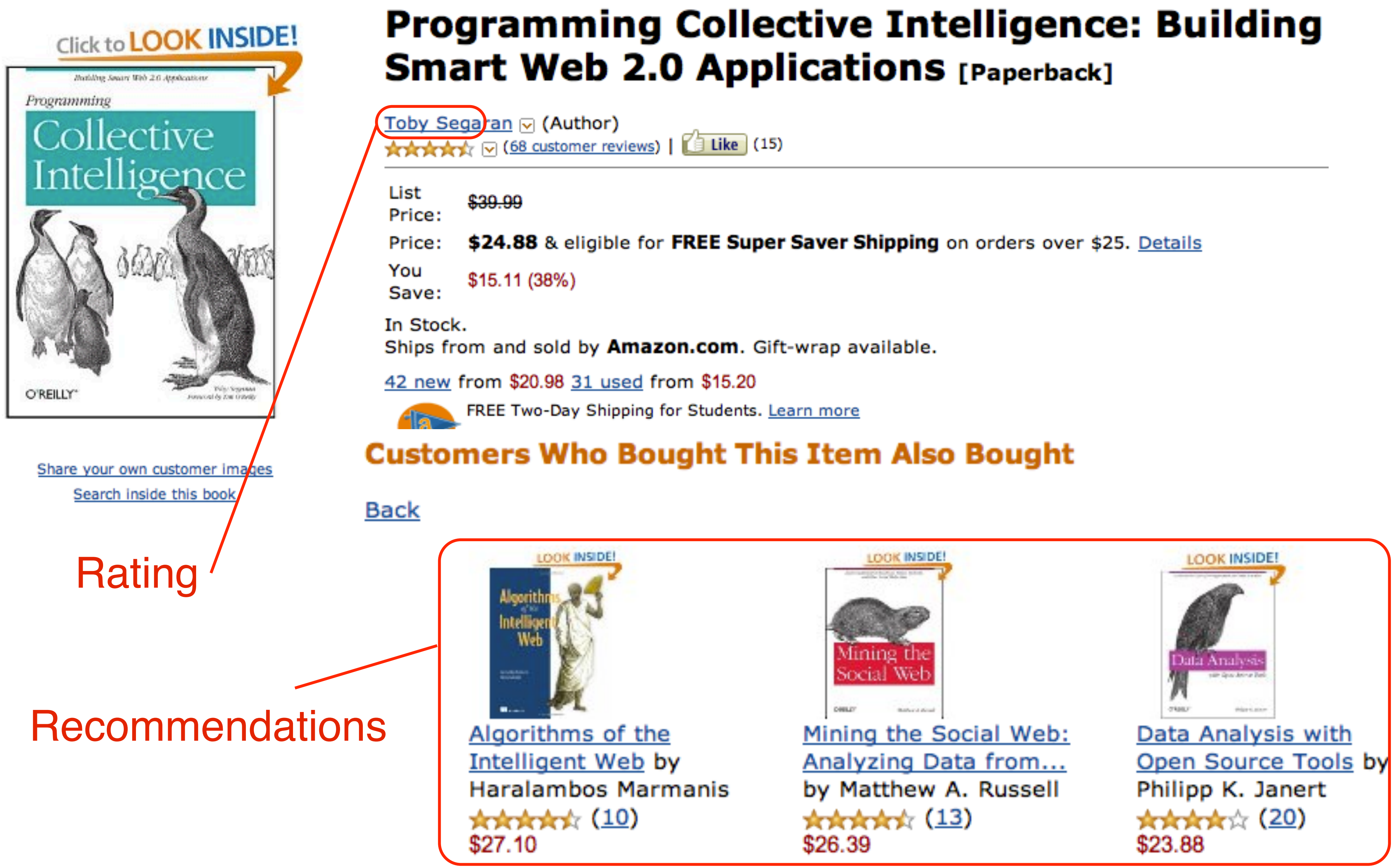
Fundamental operations

	$E[m_1]$	$E[m_1] \cdot E[m_2]$	$E[m_1]^{m_2}$
Alg. 1: Basic scheme	n-1	n-1	n-1
Alg. 2: Pre-computation	0	2(k-1)	n-1
Alg. 3: k-Nearest neighbor	k-1	k-1	k-1
Alg. 2 with Alg. 3	0	2(k-1)	k-1

Basic algorithm

- Organizations A and B
- B wish to predict rating values
  - Step 1 : B **sends** target ID and **ciphertexts** to A.
  - Step 2 : A and B **compute** similarities between users and **locally predict rating**.
  - Step 3 : A **sends the prediction of rating** to B.
  - Step 4 : B decrypts the ciphertext and **aggregates A and B ratings.**

What is Recommendation?



Share your own customer images

Search inside this book

Programming Collective Intelligence: Building Smart Web 2.0 Applications [Paperback]

Toby Segaran (Author)

★★★★☆ (68 customer reviews) | Like (15)

List Price:	~~$39.99~~
Price:	**$24.88** & eligible for **FREE Super Saver Shipping** on orders over $25. Details
You Save:	$15.11 (38%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

42 new from $20.98 31 used from $15.20

FREE Two-Day Shipping for Students. Learn more

Customers Who Bought This Item Also Bought

Back

Algorithms of the Intelligent Web by Haralambos Marmanis
★★★★☆ (10)
$27.10

Mining the Social Web: Analyzing Data from... by Matthew A. Russell
★★★★☆ (13)
$26.39

Data Analysis with Open Source Tools by Philipp K. Janert
★★★★☆ (20)
$23.88

Rating

Recommendations

Requirements

amazon

A	i_1	i_2	Sim
u_1	4	3	1
u_2	1	4	0.69
u_3	1		0.13
u_4	4	3	-

ebay

B	i_3	i_4	i_5	Sim
u_1	5	1		0.12
u_2	3		4	0.29
u_3	4	5	2	0.59
u_4	*	2	2	-

Requirement 1 — **to improve accuracy**

Requirement 2 — **to improve performance**