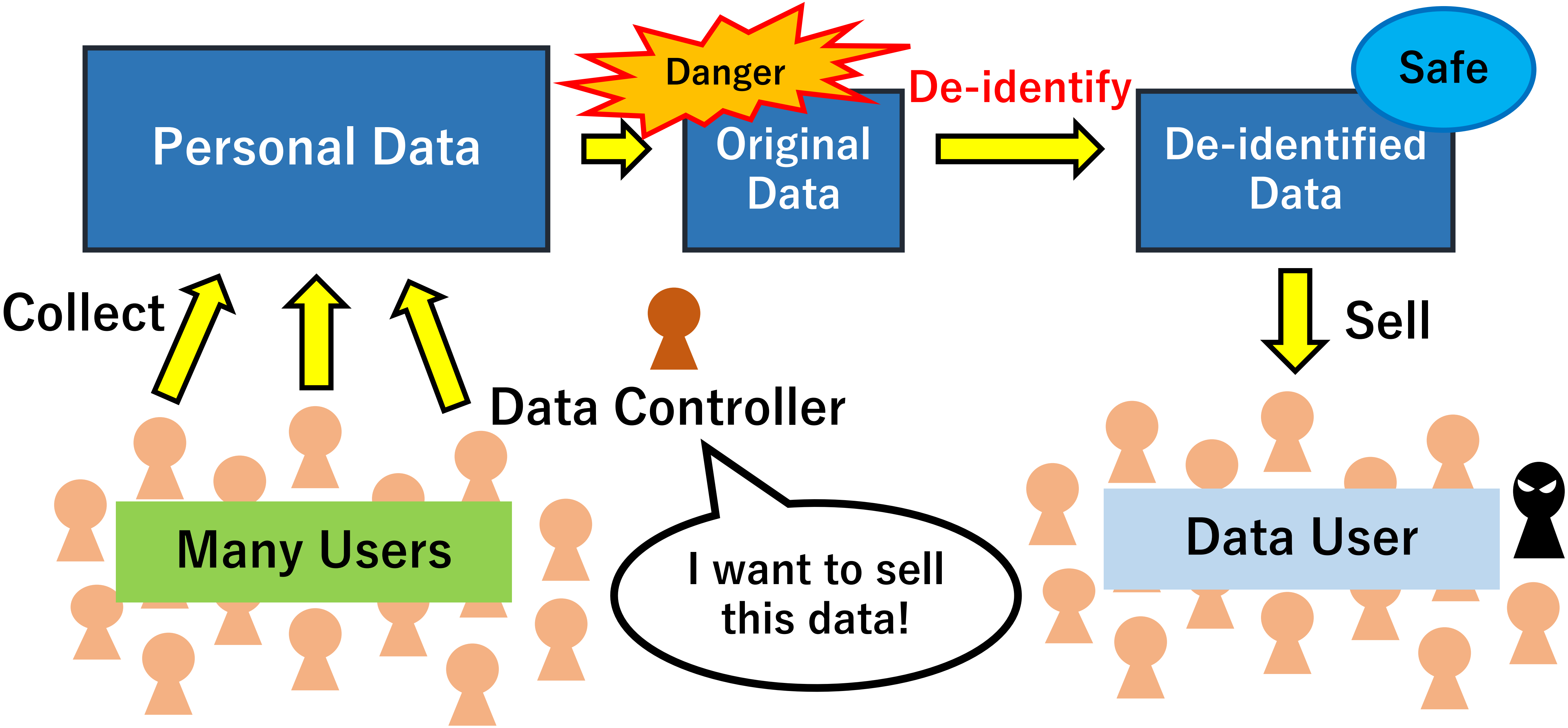
MDAI 2019

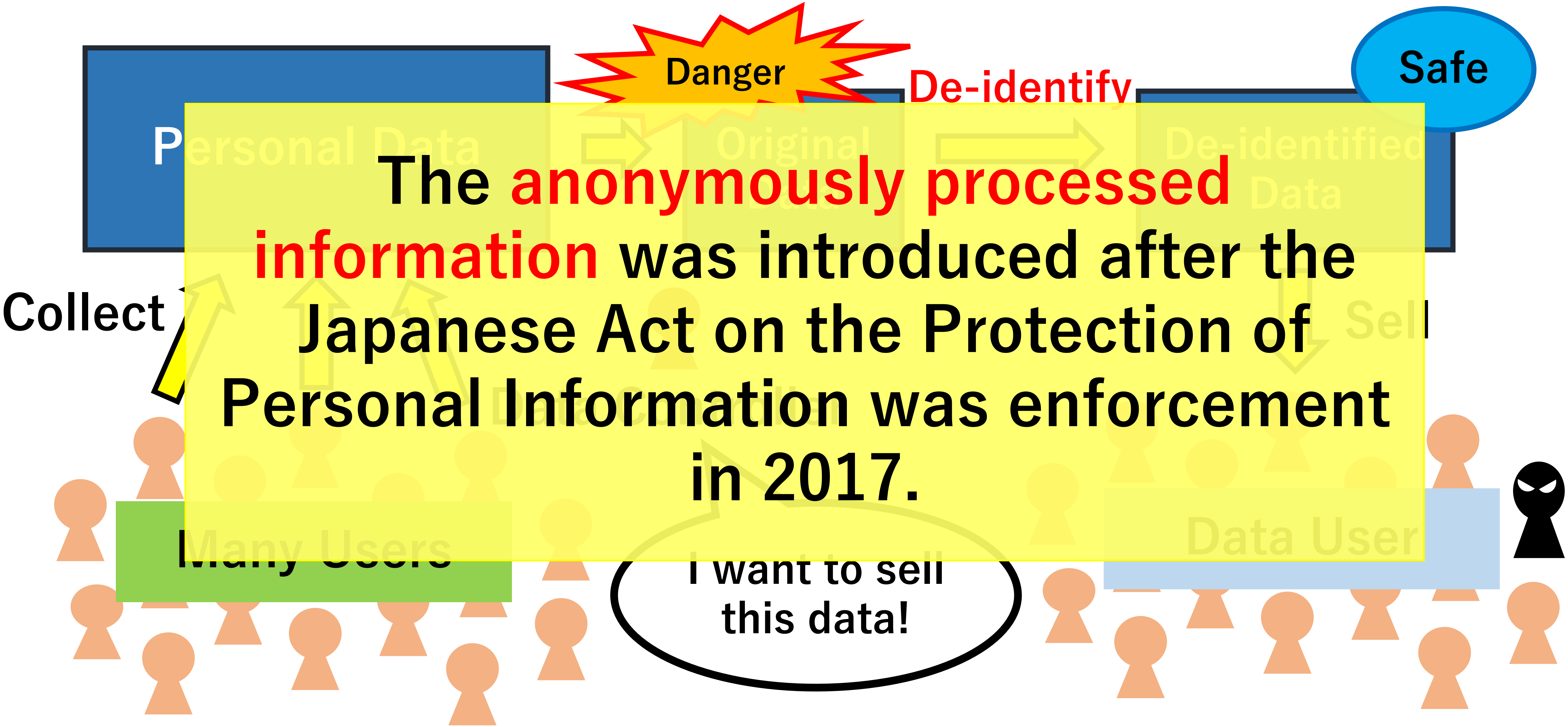
De-identification for Transaction Data Secure against Re-identification Risk Based on Payment Records

Satoshi Ito, Reo Harada, and Hiroaki Kikuchi
Meiji University

What is De-identification?



What is De-identification?



The anonymously processed information was introduced after the Japanese Act on the Protection of Personal Information was enforcement in 2017.

Record Linkage Risk from the Jaccard Coefficient

Jaccard Coefficient: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

User ID A must be replaced by pseudonym 1 !

Original Data

User ID	Goods
A	Apple
B	Apple
B	Book
C	Book

De-identified Data

Pseudonym	Goods
1	Apple
2	Book
3	Book

Attacker

$$J(A,1) = 1$$
$$J(B,1) = 1/2$$
$$J(C,1) = 0$$

How to Prevent Data from Being Distinguished with the Jaccard Coefficient

Original Data

User ID	Goods
A	Apple
B	Apple
B	Book
C	Book

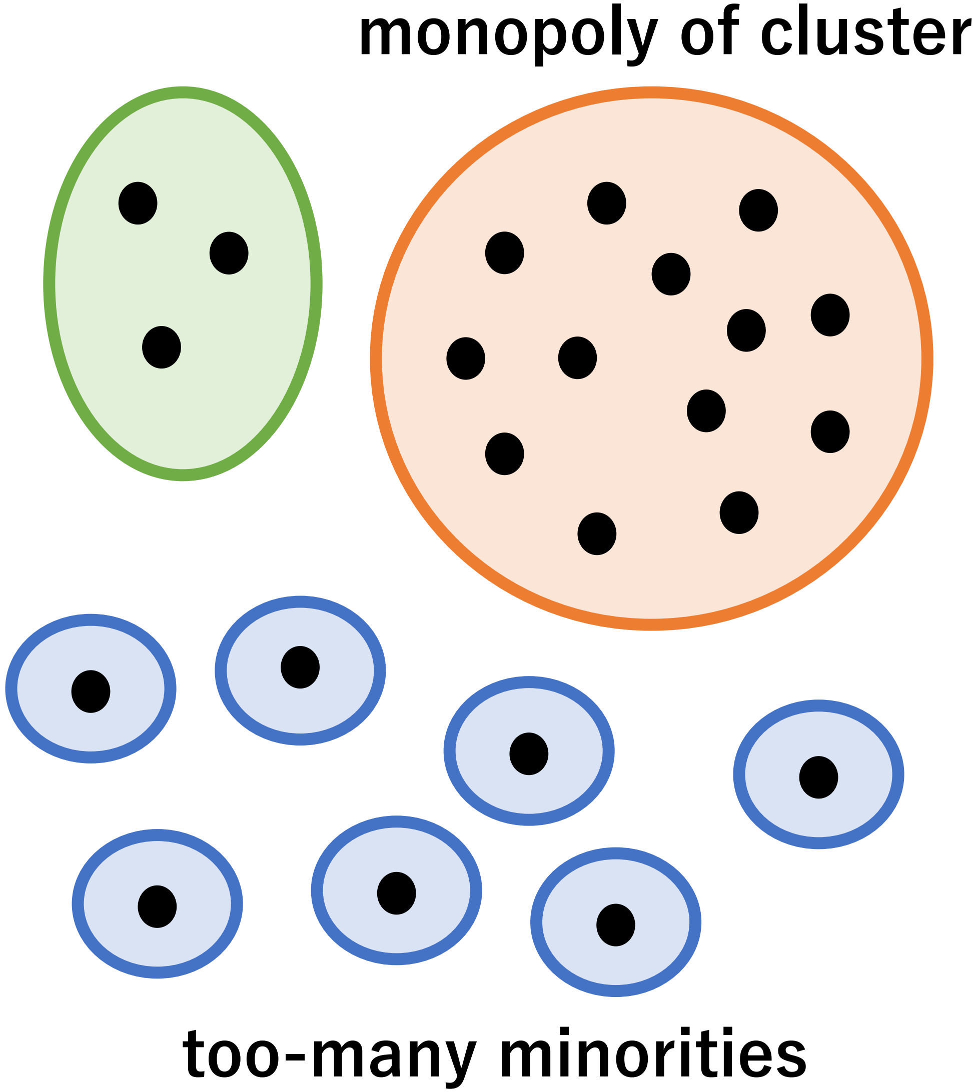
De-identified Data

Pseudonym	Goods
1	Apple
1	Book
2	Apple
2	Book
3	Apple
3	Book

Dummy Records

Re-identified rate: 1/3

Problems of classification customers



Problem 1 (the monopoly of cluster)

The utility of data will decrease because the too many dummy records are required to generalize many users belonging the large one.

Problem 2 (too-many minorities)

The privacy of data will be lost because the customers in this clusters must be identified.

How to resolve these problems

Method 1

k-means clustering based on TF-IDF

TF: Term Frequency

(Frequency of goods of customers)

IDF: Inverse document frequency

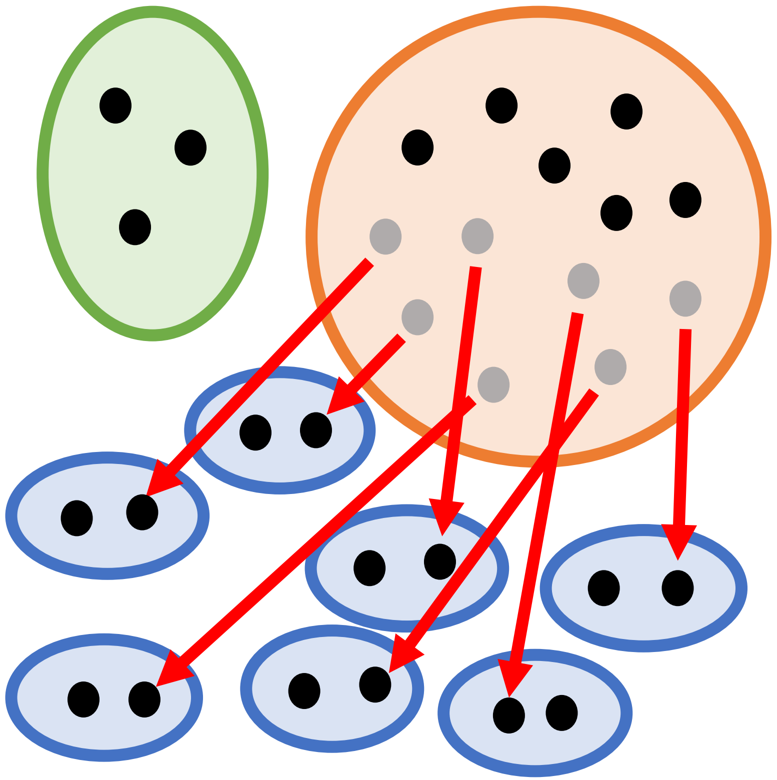
(Importance of goods)

	g_1	g_2	g_3
u_1	1	1	0
u_2	1	0	1
u_3	0	1	1
u_4	0	1	0

	g_1	g_2	g_3
u_1	0.7	0.6	0
u_2	0.7	0	0.7
u_3	0	0.6	0.7
u_4	0	1.1	0

Method 2

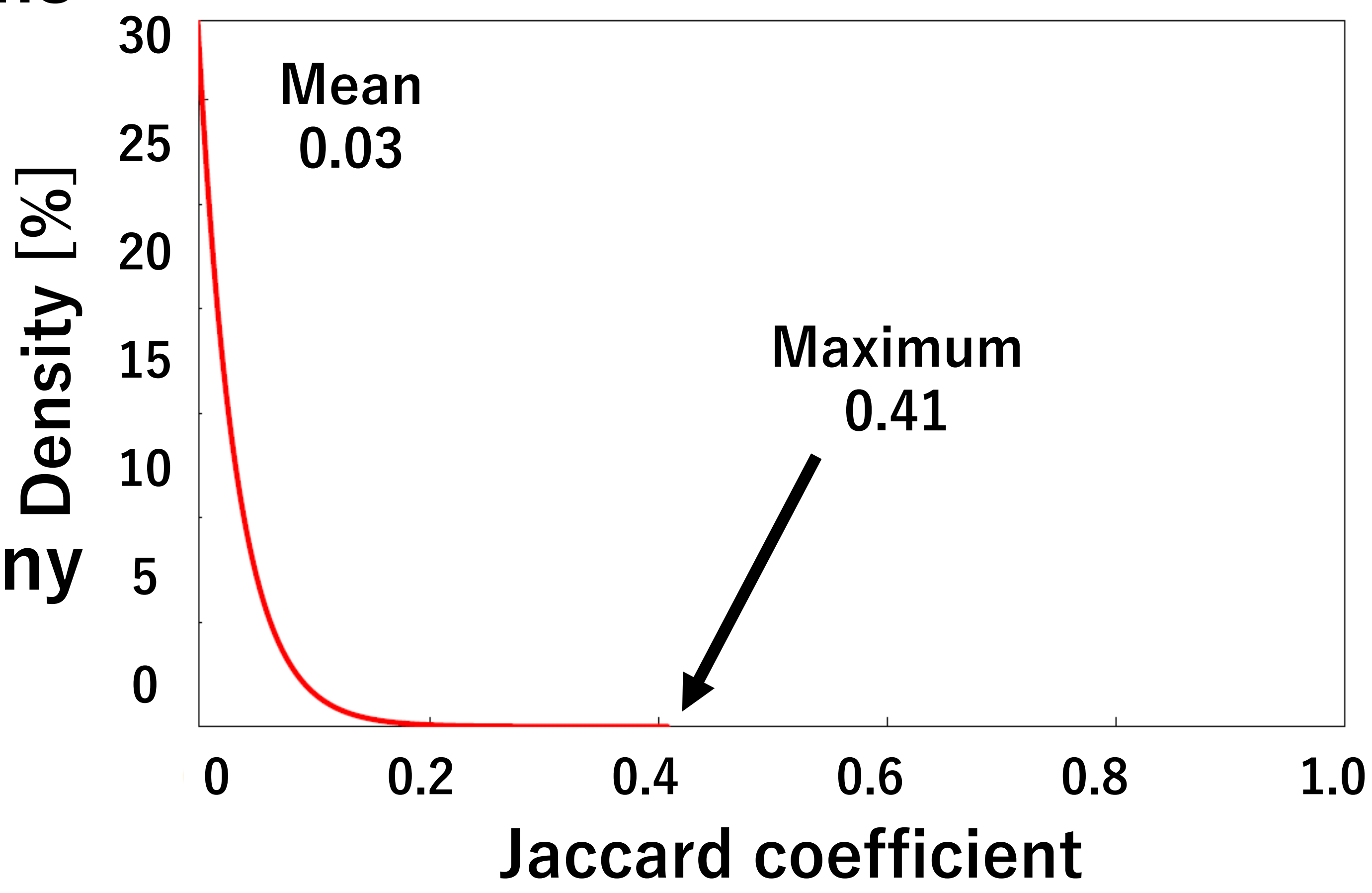
Distribution of the largest cluster



Evaluation in Online Retail Dataset (1/2)

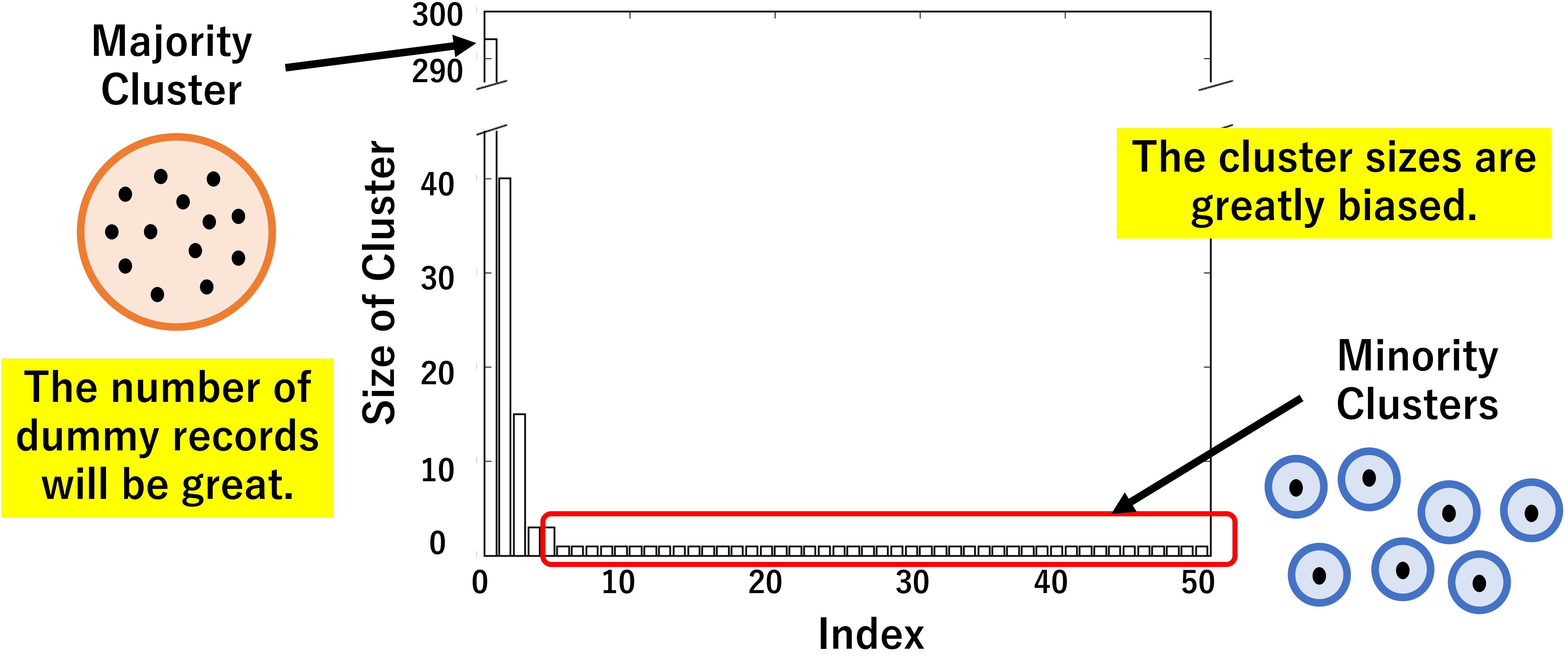
We evaluate these two problems in Online Retail Dataset.

The sets of purchased goods are quite distinct and there is great diversity in customers because this data contains many goods (2,781 goods).

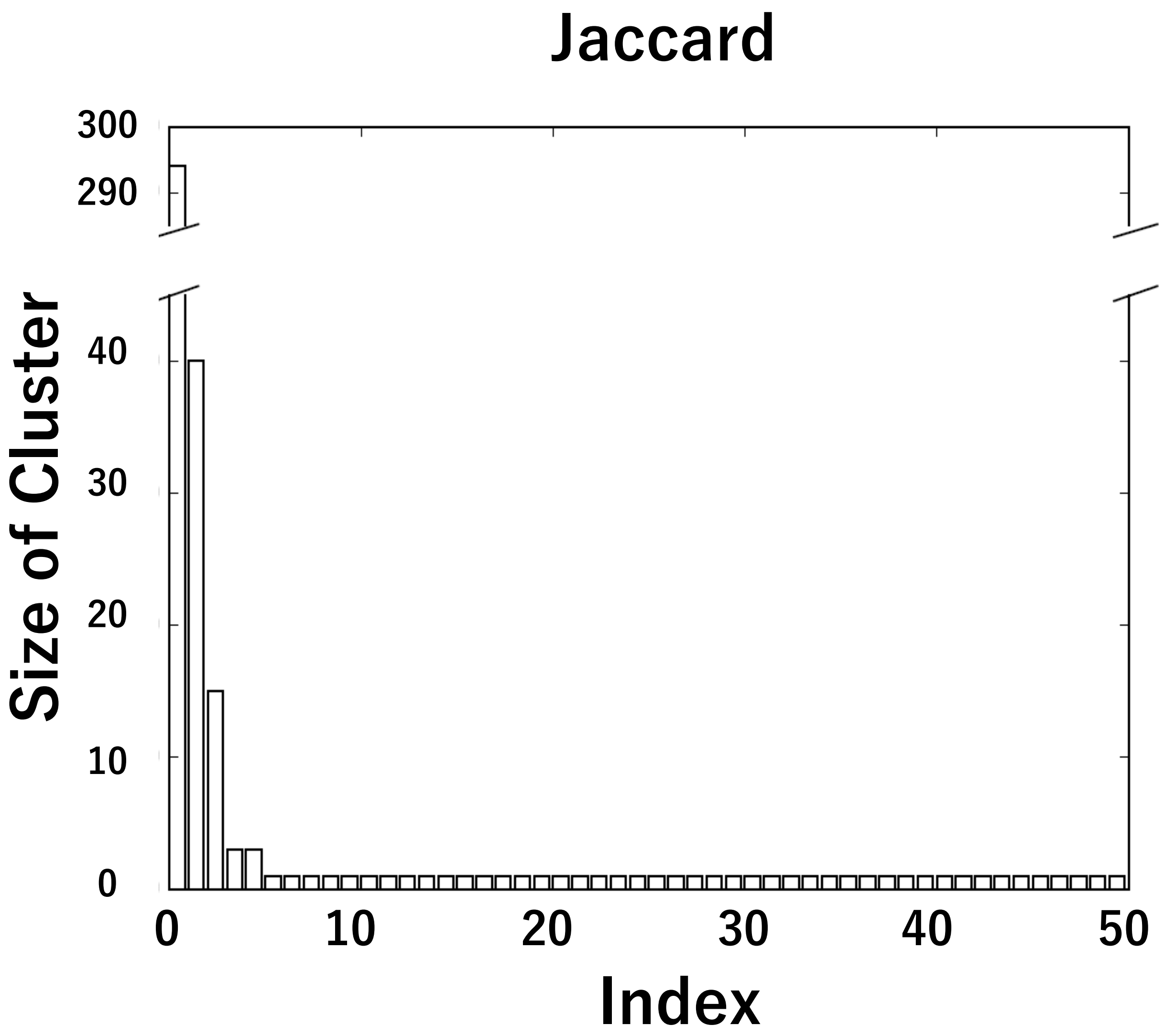


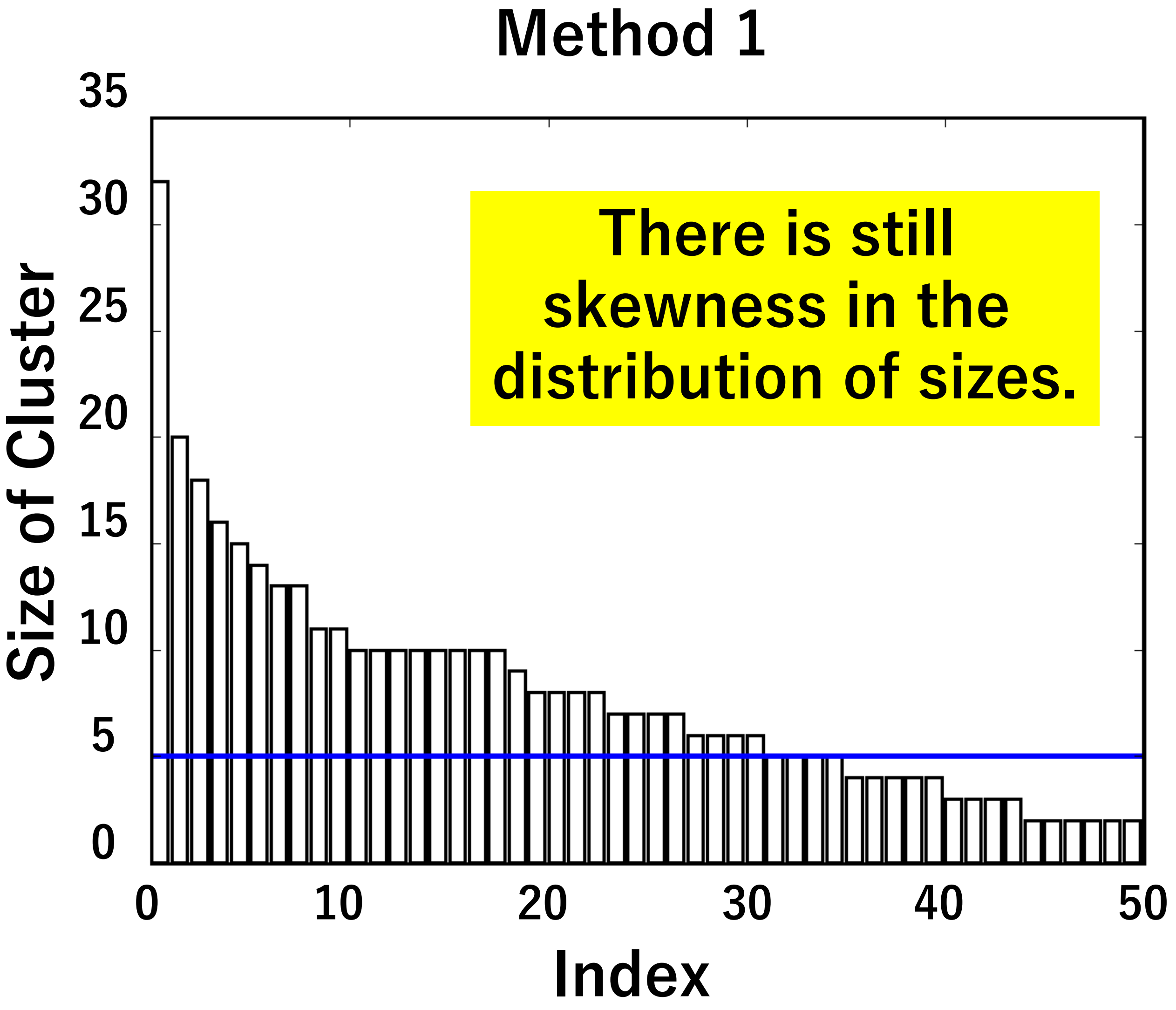
The most similar pair of customers has a similarity of only 41%.

Evaluation for Online Retail Dataset (2/2)

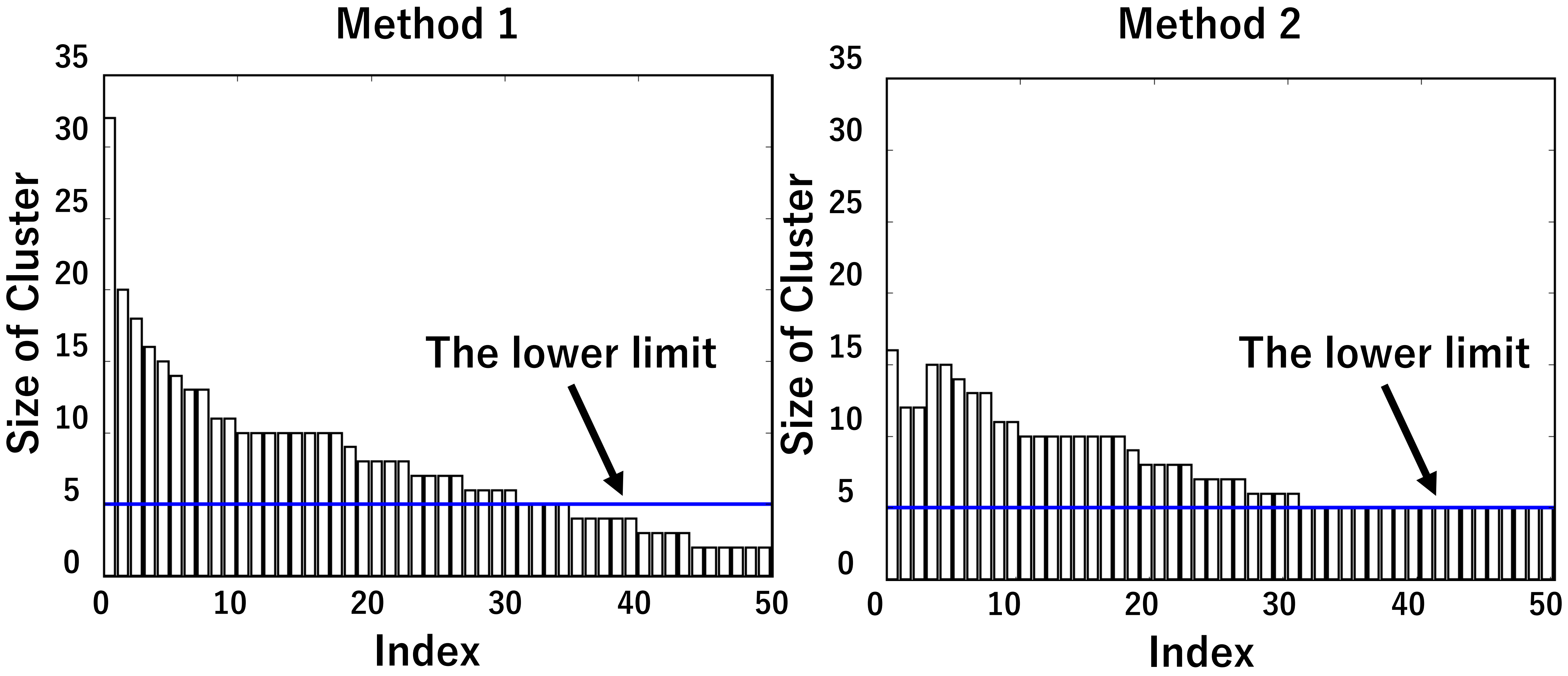


Method 1 for Online Retail Dataset





Method 2 for Online Retail Dataset



The Relationship between the lower limit and the Number of Dummy Records

	$c = 50$		$c = 100$		$c = 125$	
	Δm	Re-id	Δm	Re-id	Δm	Re-id
Method 1	182,297	0.1235	128,568	0.2488	97,581	0.3120
$s_{min} = 2$	183,902	0.1223	99,228	0.2475	60,492	0.3105
$s_{min} = 3$	175,449	0.1222	68,357	0.2480	46,101	0.3102
$s_{min} = 4$	162,474	0.1218	59,374	0.2465		
$s_{min} = 8$	125,798	0.1218				

Method 2

In general, we improve the utility of de-identified data as limit size increase.

The Relationship between the lower limit and the Number of Dummy Records

	$c = 50$		$c = 100$		$c = 125$	
	Δm	Re-id	Δm	Re-id	Δm	Re-id
Method 1	182,297	0.1235	128,568	0.2488	97,581	0.3120
$s_{min} = 2$	183,902	0.1223	99,228	0.2475	60,492	0.3105
$s_{min} = 3$	175,449	0.1222	68,357	0.2480	46,101	0.3102
$s_{min} = 4$	162,474	0.1218	59,374	0.2465		
$s_{min} = 8$	125,798	0.1218				

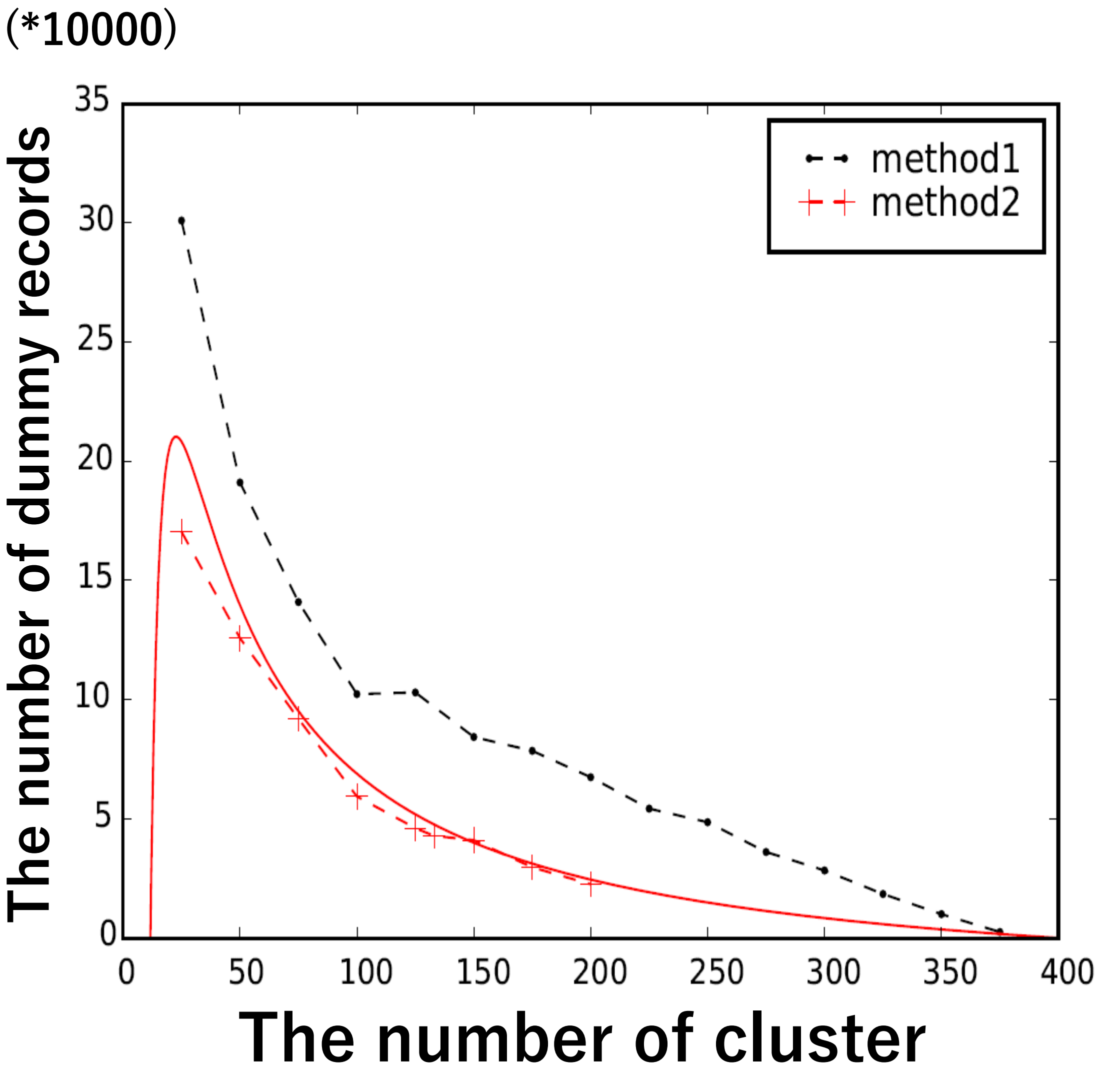
Method 2

You can see the re-identification rates are almost stable in the column.

Theory in the Number of Dummy Records

n : the number of customers
c : the number of clusters
b : the mean number of goods that a customer purchases in a year
h : mean size of the intersection of the two sets of goods purchased by distinct customers
Δm : the number of dummy records

$$E(\Delta m) = -\frac{hn^3}{2c^2} + \left(b + \frac{h}{2}\right)\frac{n^2}{c} - bc$$



Conclusions

- **We revealed the risk of purchasing goods of customers and proposed a new de-identification method by reducing additional dummy records.**

- **We have demonstrated that our proposed algorithm reduces the number of dummy records as far as restricted size of clusters.**

- **We estimated the expected value of the number of dummy records.**