垂直分割における通信効率の良い一致度の秘匿分散計算

青木 良樹† 菊池 浩明†
寺田 雅之‡ 石井 一彦* 関野 公彦*

† 東海大学 ‡ NTTドコモ先進技術研究所 * NTTドコモサービス&ソリューション開発部

論文の訂正

誤

3.2 提案方式：一致度の推定

k 人の類似ユーザを作り，そのユーザと
組の一致度を求める．

次に，一致度の推移性を利用し，任意の n^2 組の一致度を推定する．必要な通信コストは $nk = O(n)$ であり，効率が良いが，推定される一致度は誤差を含む．

Lathia らの提案した擬似ユーザを利用して $O(n^2)$ を n に削減する方法を提案する．まず，Algorithm?? をつかって，三組織間で擬似ユー

Algorithm 1 一致度の予測方法

入力: $\mathcal{A}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{A}})$，$\mathcal{B}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{B}})$，$\mathcal{C}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{C}})$．

出力: 擬似ユーザと各ユーザの C,T,D,N $\hat{C}_{u,c}, \hat{D}_{u,c}, \hat{T}_{u,c}, N$．

4. 式 (3, 4, 5) の区間推定を行い，全アイテム間 u, v について $[\underline{C_{uv,c_1}}, \overline{C_{uv,c_1}}], \ldots, [\underline{C_{uv,c_k}}, \overline{C_{uv,c_k}}]$ を求め，その集約区間 $\underline{C_{uv,*}} = \vee_k \underline{C_{uv,c_k}}$, $\overline{C_{uv,*}} = \vee_k \overline{C_{uv,c_k}}$ を求める．D と T についても同様に算出する．

正

3.2 提案方式：一致度の推定

k 人の類似ユーザを作り，そのユーザと
組の一致度を求める．

次に，一致度の推移性を利用し，任意の n^2 組の一致度を推定する．必要な通信コストは $nk = O(n)$ であり，効率が良いが，推定される一致度は誤差を含む．

Lathia らの提案した擬似ユーザを利用して $O(n^2)$ を $O(n)$ に削減する方法を提案する．まず，Algorithm2 をつかって，三組織間で擬似

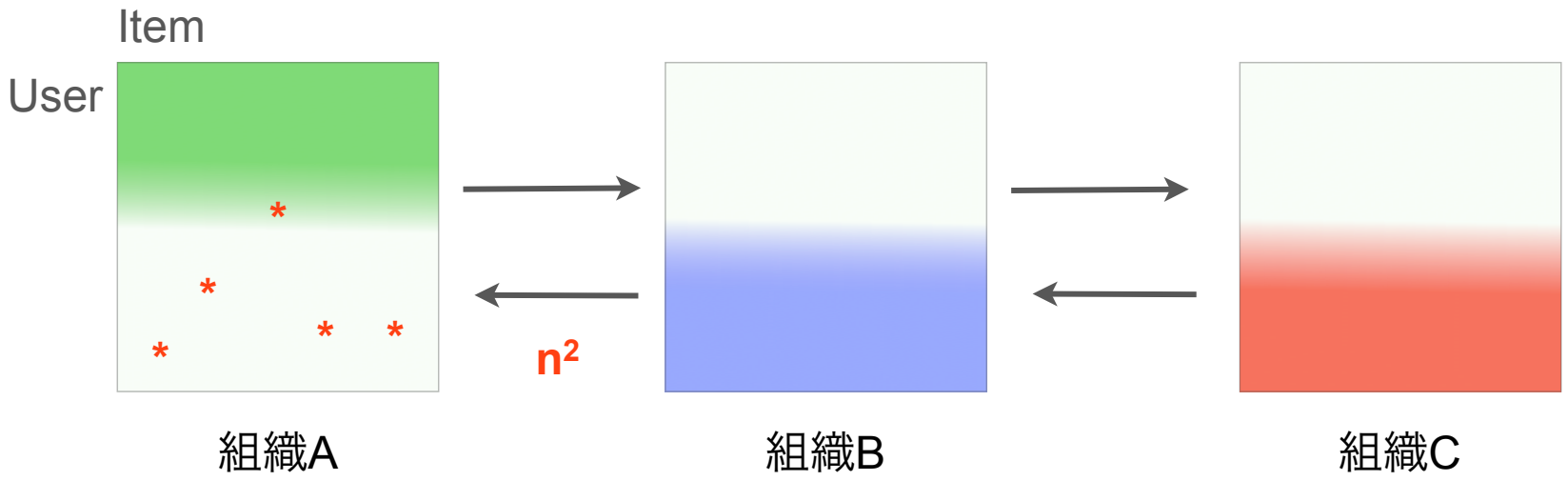
Algorithm 1 一致度区間の予測方法

入力: $\mathcal{A}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{A}})$，$\mathcal{B}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{B}})$，$\mathcal{C}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{C}})$．

出力: 集約区間 $[\underline{C_{uv,*}}, \overline{C_{uv,*}}]$，$[\underline{D_{uv,*}}, \overline{D_{uv,*}}]$, $[\underline{T_{uv,*}}, \overline{T_{uv,*}}]$．

4. 式 (3, 4, 5) の区間推定を行い，全アイテム間 u, v について $[\underline{C_{uv,c_1}}, \overline{C_{uv,c_1}}], \ldots, [\underline{C_{uv,c_k}}, \overline{C_{uv,c_k}}]$ を求め，その集約区間 $\underline{C_{uv,*}} = \vee_k \underline{C_{uv,c_k}}$, $\overline{C_{uv,*}} = \wedge_k \overline{C_{uv,c_k}}$ を求める．D と T についても同様に算出する．

垂直分割DM



目的

B,Cの評価値を基に，類似度を正確に予測する

それぞれの組織の評価値は秘匿

問題点

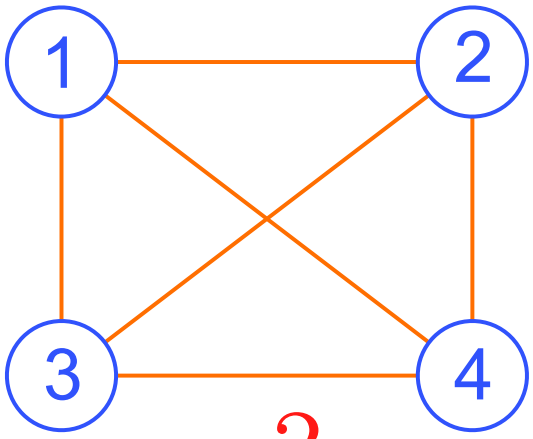
全ユーザ間の組み合わせ$O(n^2)$の情報交換

既存研究

	手法	通信コスト
[VC03]	k-means	$O(n^2)$
[BVK11]	Slope One+CF	$O(n^2)$
本発表	**Concordant**	**O(kn)**

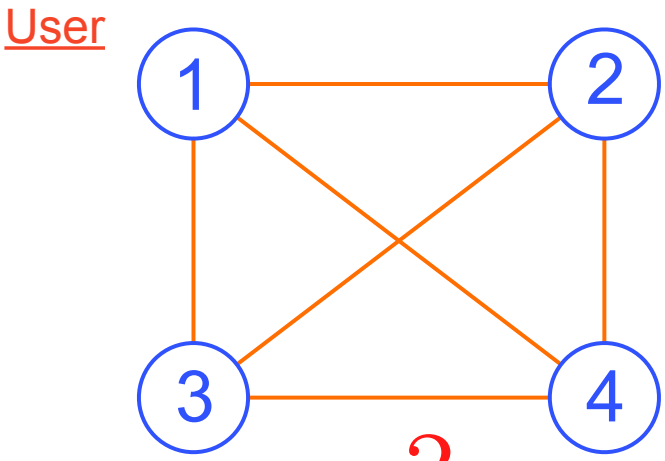
本アプローチの新規性

User

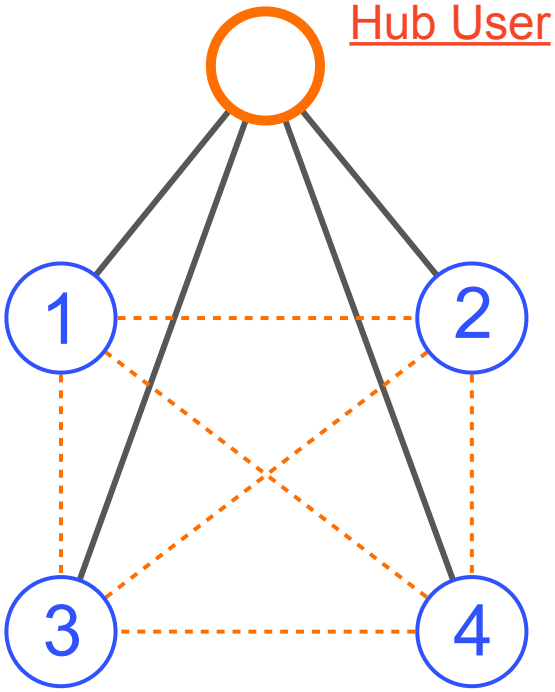


約 n^2 通り

本アプローチの新規性



約 n^2 通り



n

＋

Concordant(一致度)に基づく

類似度推定[Lathia 2007]

一致度の定義 （Somers' d）

一致 (**C**oncordant)

不一致(**D**iscordant)

同順位 (**T**ied)

	i_1	i_2	i_3	i_4	i_5
u_1	4	2	5	1	5
	+	-	+	-	+
u_2	1	2	3		4
	-	-	+	/	+
$u_{1,2}$	**d**	**c**	**c**	**t**	**c**

平均値より
＋？－？

一致度の定義 （Somers' d）

一致 (**C**oncordant)

不一致(**D**iscordant)

同順位 (**T**ied)

	i_1	i_2	i_3	i_4	i_5
u_1	4	2	5	1	5
	+	-	+	-	+
u_2	1	2	3		4
	-	-	+	/	+
$u_{1,2}$	d	c	c	t	c

平均値より
＋？－？

$$C_{1,2} = 3$$

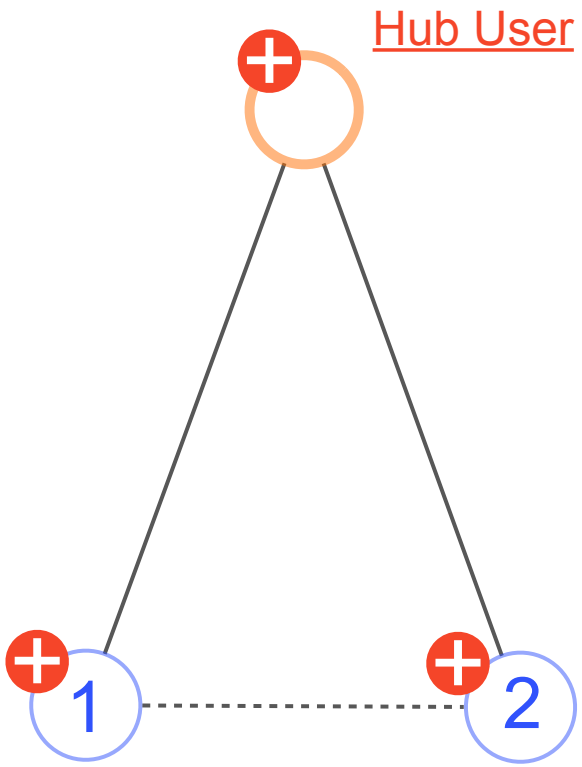
$$D_{1,2} = 1$$

$$T_{1,2} = 1$$

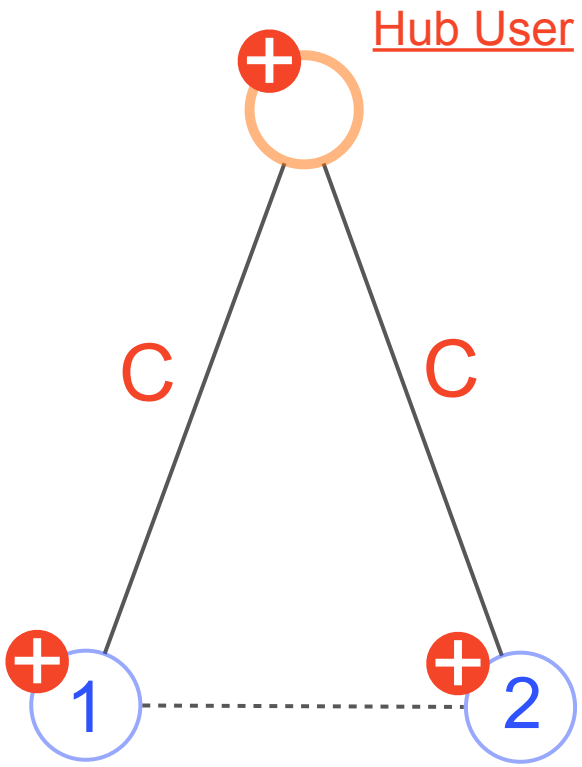
$$d_{a,b} = \frac{C - D}{N - T}$$

$$d_{1,2} = \frac{3 - 1}{5 - 1} = 0.5$$

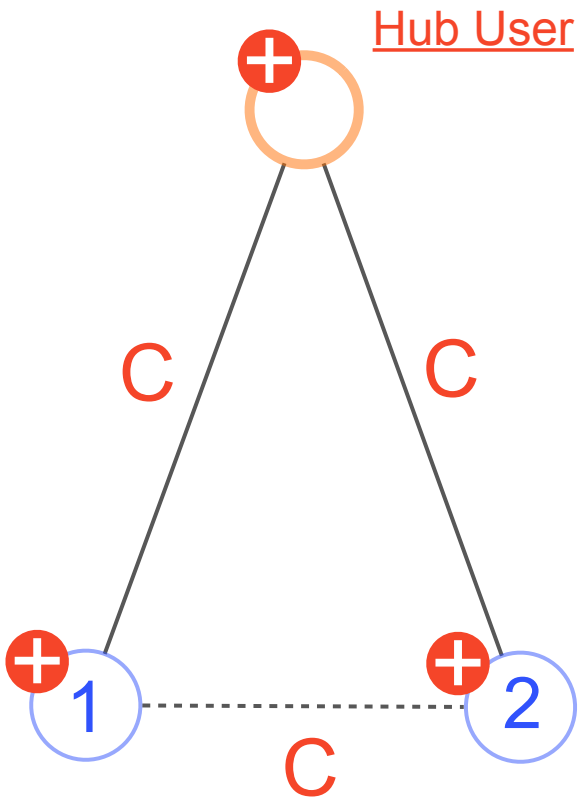
一致度の性質



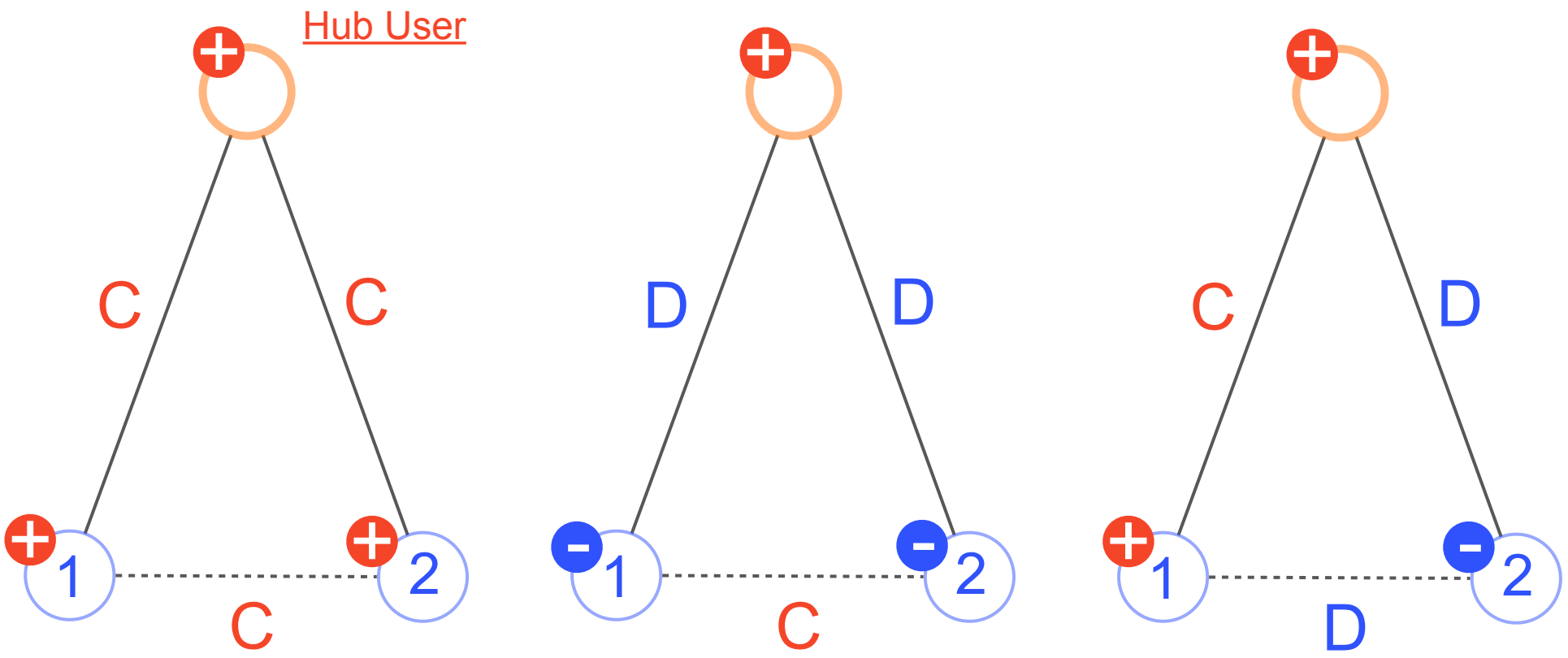
一致度の性質



一致度の性質



一致度の性質



一致度の推定 （Concordantの推定） [Lathia07]

最悪のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	c	c	c	c	d	d	d	t
$u_{2,hub}$	d	d	d	d	c	c	t	t
$u_{1,2}$	d	d	d	d	d	d	t	t

下限

$$\underline{C_{12}} = 0$$

一致度の推定 （Concordantの推定） [Lathia07]

最悪のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	c	c	c	c	d	d	d	t
$u_{2,hub}$	d	d	d	d	c	c	t	t
$u_{1,2}$	d	d	d	d	d	d	t	t

下限

$$\underline{C_{12}} = 0$$

最良のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	c	c	c	c	d	d	d	t
$u_{2,hub}$	c	c	t	t	d	d	d	d
$u_{1,2}$	c	c	t	t	c	c	c	t

上限

$$\overline{C_{12}} = 5$$

一致度の推定 （Concordantの推定） [Lathia07]

最悪のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	c	c	c	c	d	d	d	t
$u_{2,hub}$	d	d	d	d	c	c	t	t
$u_{1,2}$	d	d	d	d	d	d	t	t

下限
$\underline{C_{12}} = 0$

推定値

最良のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	c	c	c	c	d	d	d	t
$u_{2,hub}$	c	c	t	t	d	d	d	d
$u_{1,2}$	c	c	t	t	c	c	c	t

上限
$\overline{C_{12}} = 5$

一致度の推定 （Tied）[Lathia07]

最悪のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	t	t	t	t	c	c	d	d
$u_{2,hub}$	t	t	t	t	t	c	d	d
$u_{1,2}$	t	t	t	t	t	c	c	c

下限
$\underline{T_{12}} = 5$

最良のケース	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
$u_{1,hub}$	t	t	t	t	c	c	d	d
$u_{2,hub}$	d	d	c	t	t	t	t	t
$u_{1,2}$	t	t	t	t	t	t	t	t

上限
$\overline{T_{12}} = 8$

$$\max(T_{uc}, T_{vc}) \leq T_{uv} \leq \min(T_{uc} + T_{vc}, N)$$

一致度の推定式

- 一致 (Concordant)

$$\max(C_{uc} + C_{vc} - N, 0) + \max(D_{uc} + D_{vc} - N, 0)$$
$$\leq C_{uv} \leq \min(C_{uc}, C_{vc}) + \min(D_{uc}, D_{vc})$$

- 不一致(Discordant)

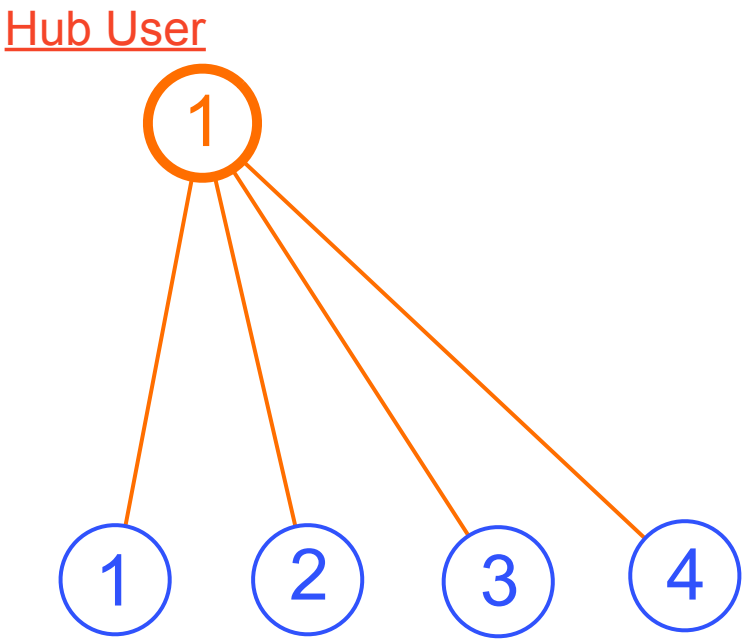
$$\max(N - (\overline{C_{uv}} + \overline{T_{uv}}), 0) \leq D_{uv} \leq N - (\underline{C_{uv}} + \underline{T_{uv}})$$

- 同順位 (Tied)

$$\max(T_{uc}, T_{vc}) \leq T_{uv} \leq \min(T_{uc} + T_{vc}, N)$$

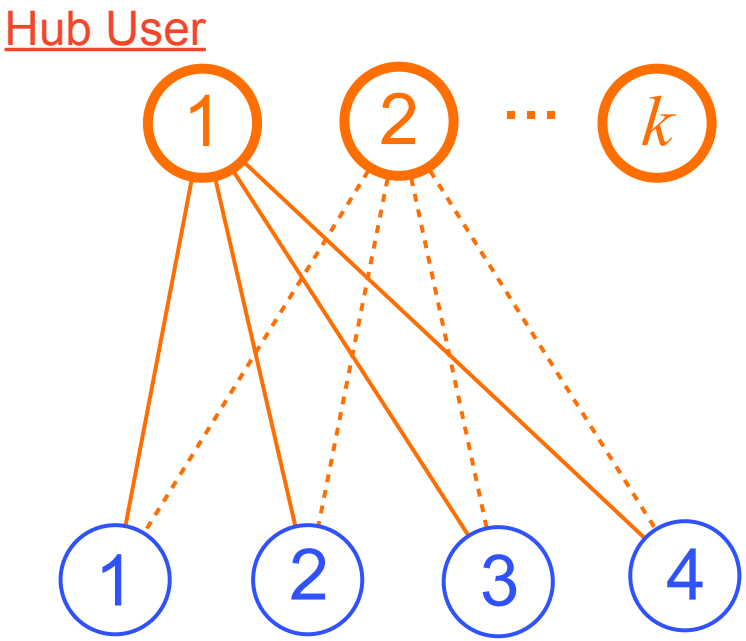
提案方式：区間の集約方式

Hubユーザ c_j を独立に k 人作成



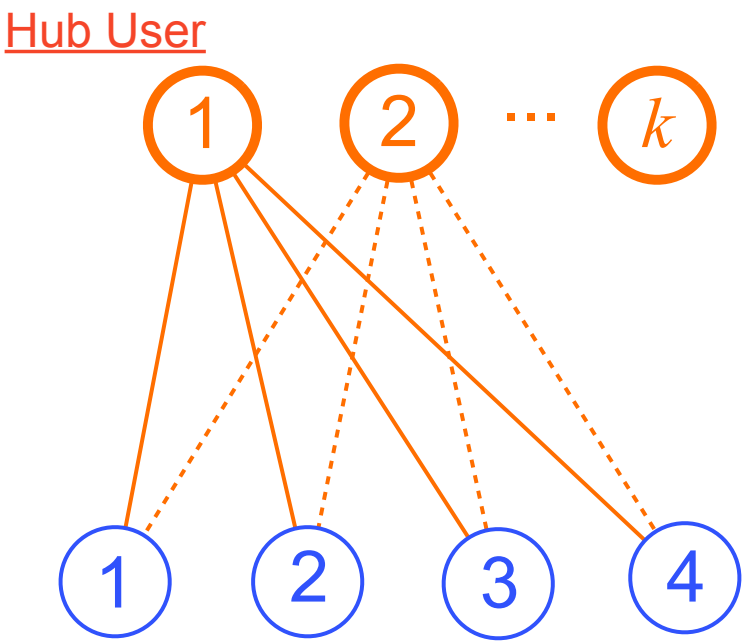
提案方式：区間の集約方式

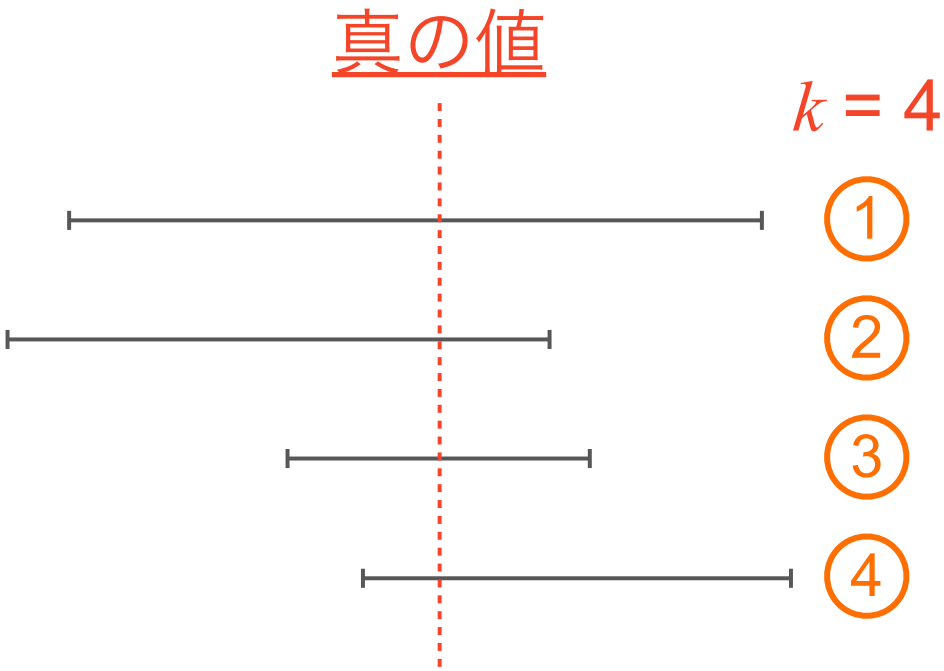
Hubユーザ c_j を独立に k 人作成



提案方式：区間の集約方式

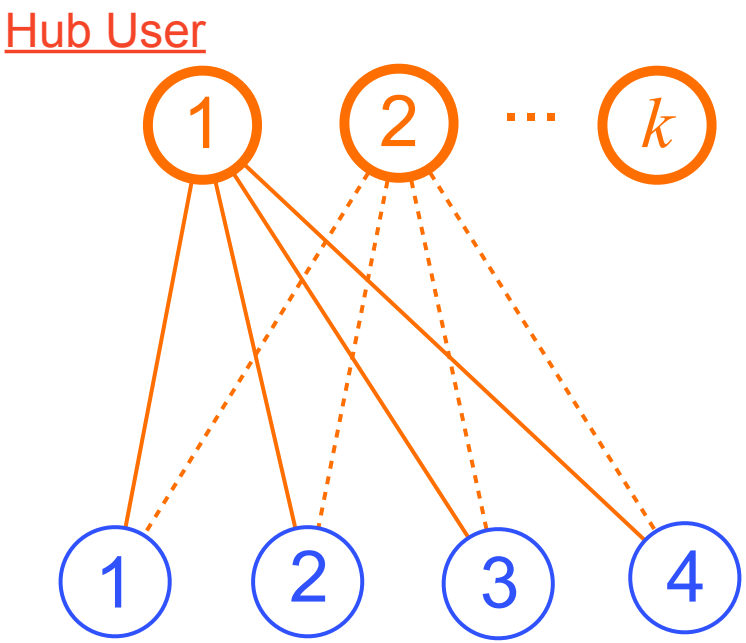
Hubユーザ c_j を独立に k 人作成

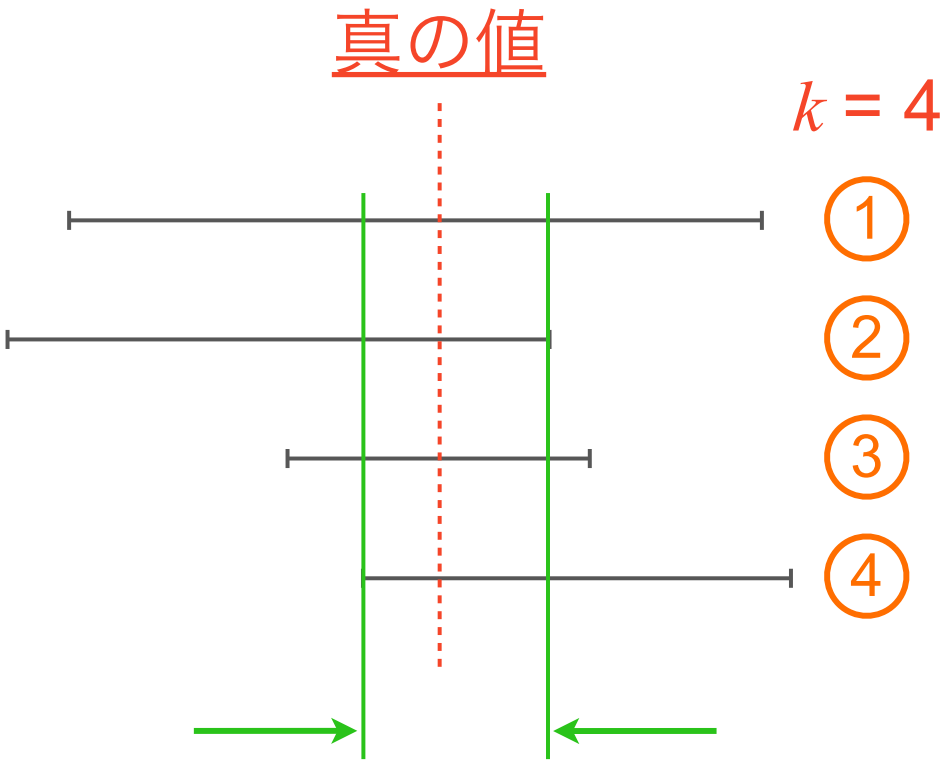




提案方式：区間の集約方式

Hubユーザ c_j を独立に k 人作成





提案通信プロトコル (Algorithm 1)

Step1

	組織A				組織B				組織C	
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	2		4	3	4	1	2	3	2
u_2	3	5	1		2		3	4	2	3
c_1	1	2	4	2	4	5	3	2	1	2

提案通信プロトコル (Algorithm 1)

Step1

	組織A				組織B				組織C	
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	2		4	3	4	1	2	3	2
u_2	3	5	1		2		3	4	2	3
c_1	1	2	4	2	4	5	3	2	1	2

Step2

	C	D	T	N	C	D	T	N	C	D	T	N
u_1:c_1	2	1	1	4	4	0	0	4	0	2	0	2
u_2:c_1	0	2	2	4	0	2	2	4	2	0	0	2

提案通信プロトコル (Algorithm 1)

Step1

	組織A				組織B				組織C	
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	2		4	3	4	1	2	3	2
u_2	3	5	1		2		3	4	2	3
c_1	1	2	4	2	4	5	3	2	1	2

Step2

	C	D	T	N	C	D	T	N	C	D	T	N
u_1:c_1	2	1	1	4	4	0	0	4	0	2	0	2
u_2:c_1	0	2	2	4	0	2	2	4	2	0	0	2

Step3

	C	D	T	N
u_1:c_1	6	3	1	10
u_2:c_1	2	4	4	10

Alg. 2 秘匿総和プロトコルで共有

提案通信プロトコル (Algorithm 1)

Step1

	組織A				組織B				組織C	
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	2		4	3	4	1	2	3	2
u_2	3	5	1		2		3	4	2	3
c_1	1	2	4	2	4	5	3	2	1	2

Step2

	C	D	T	N	C	D	T	N	C	D	T	N
u_1:c_1	2	1	1	4	4	0	0	4	0	2	0	2
u_2:c_1	0	2	2	4	0	2	2	4	2	0	0	2

Step3

	C	D	T	N
u_1:c_1	6	3	1	10
u_2:c_1	2	4	4	10

Alg. 2 秘匿総和プロトコルで共有

Step4

$$[\underline{C_{12,c_1}}, \overline{C_{12,c_1}}] = [0, 5] \qquad [\underline{T_{12,c_1}}, \overline{T_{12,c_1}}] = [4, 5]$$

$$[\underline{D_{12,c_1}}, \overline{D_{12,c_1}}] = [0, 6]$$

区間推定，出力

実験

目的

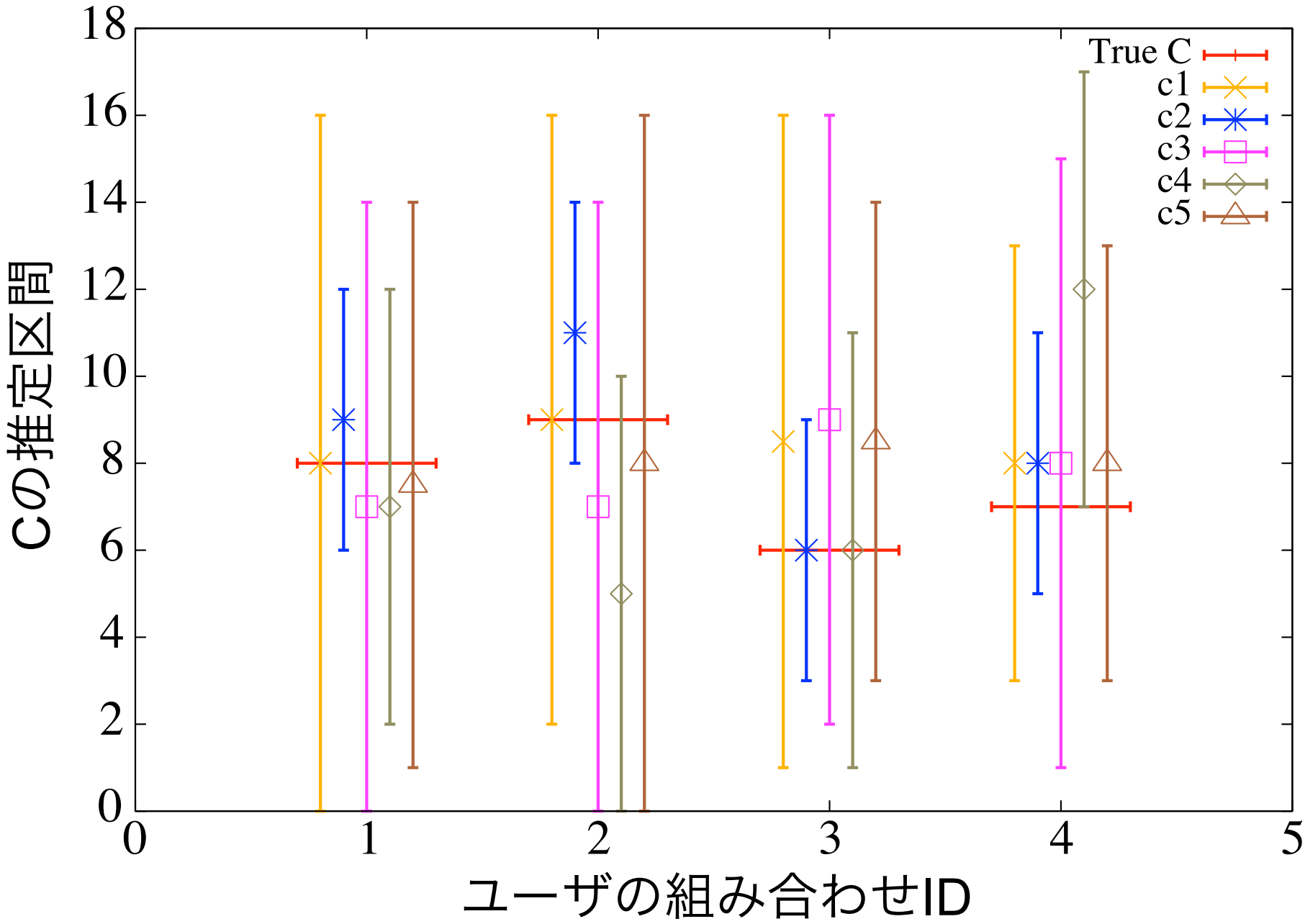
- Hubユーザ数 k の増加に伴う区間の変化
- 予測した類似度と真の類似度の差

方法

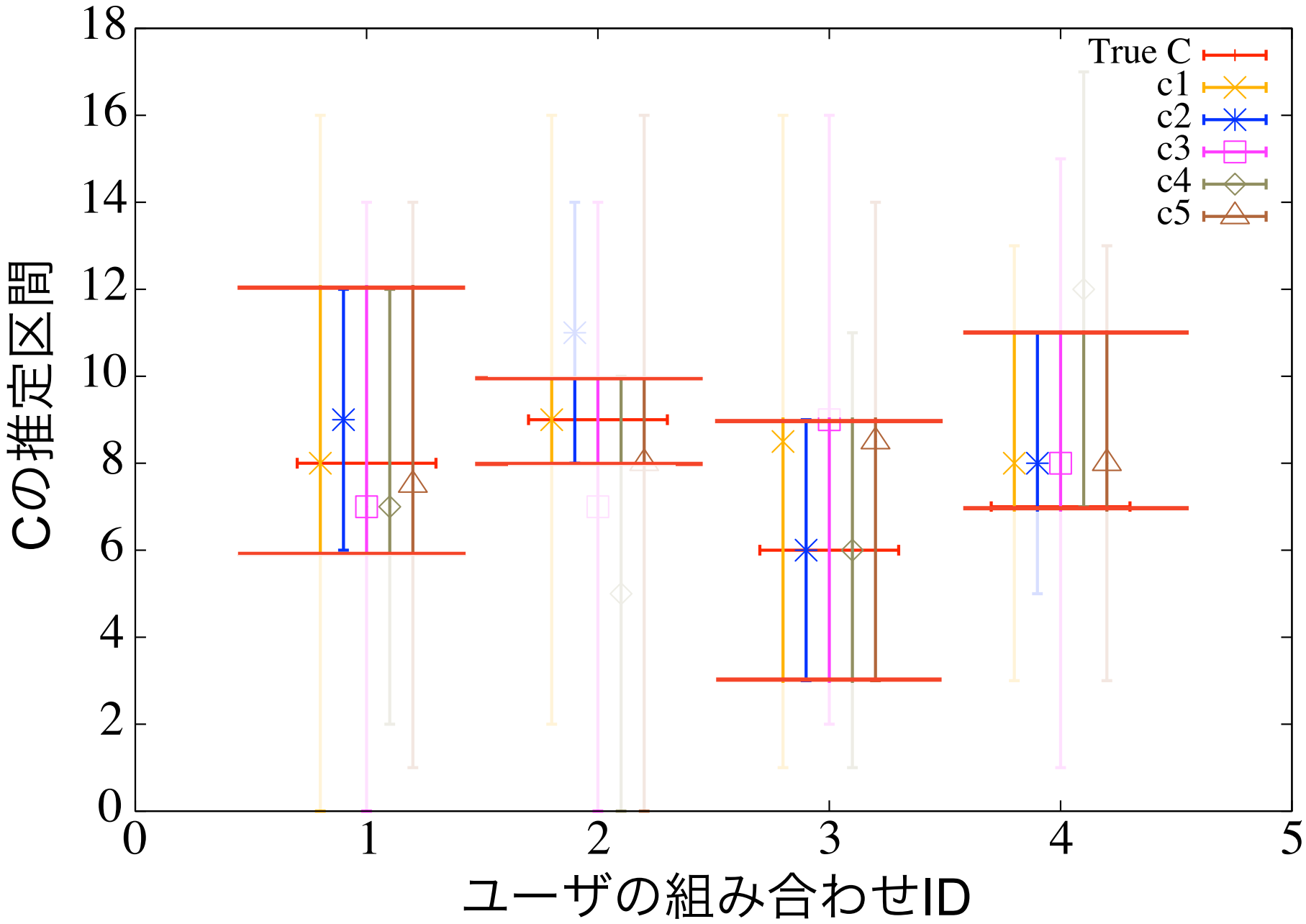
ユーザ数 6, アイテム数 17, 評価数 102, のランダムデータセットを使い提案方式で区間を推定.

- 結果1：擬似ユーザによる推定区間の変化
- 結果2：擬似ユーザ数 k の変化による区間の変化
- 結果3：真の類似度と推定した類似度の分布

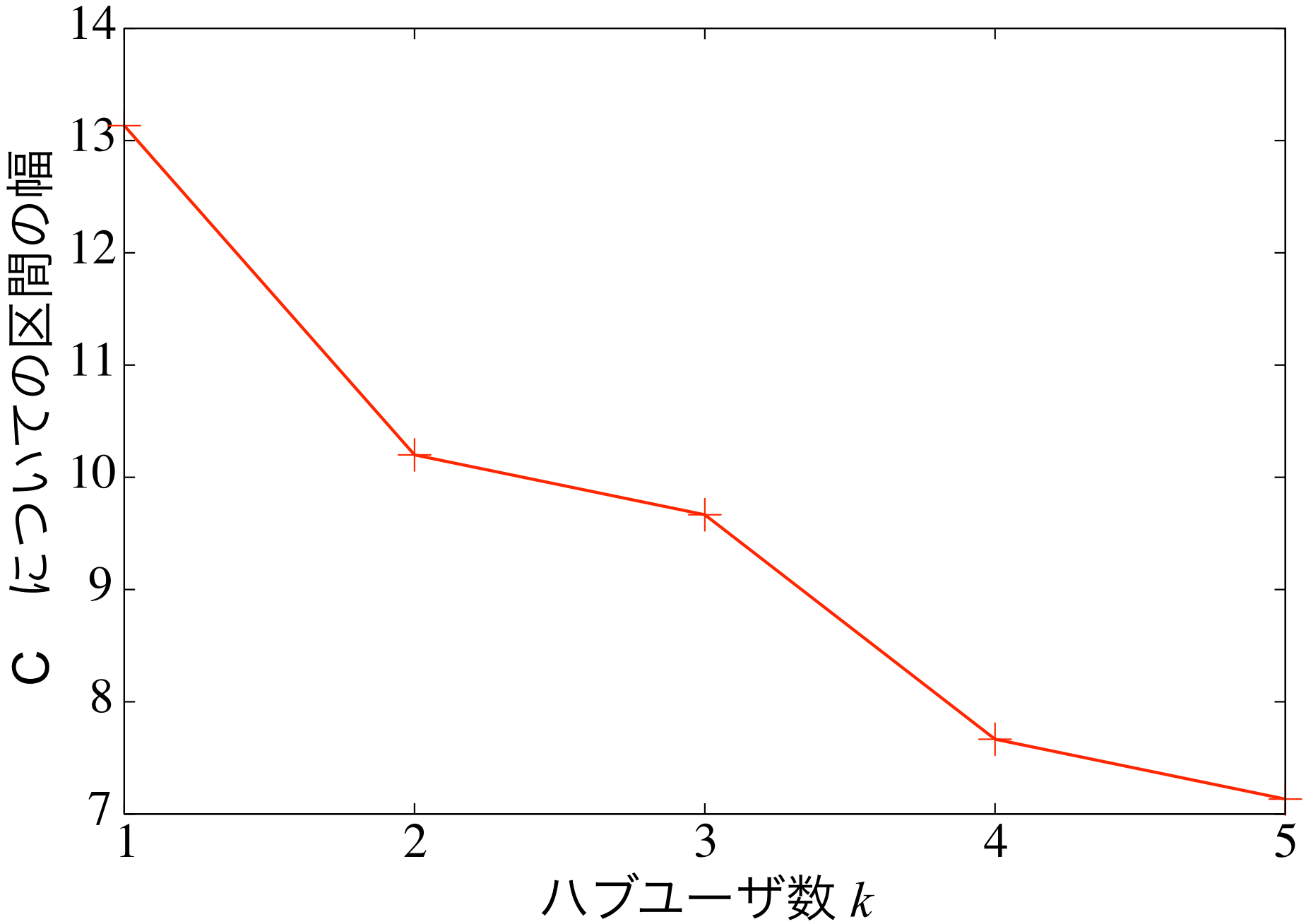
結果1：擬似ユーザによる推定区間の変化



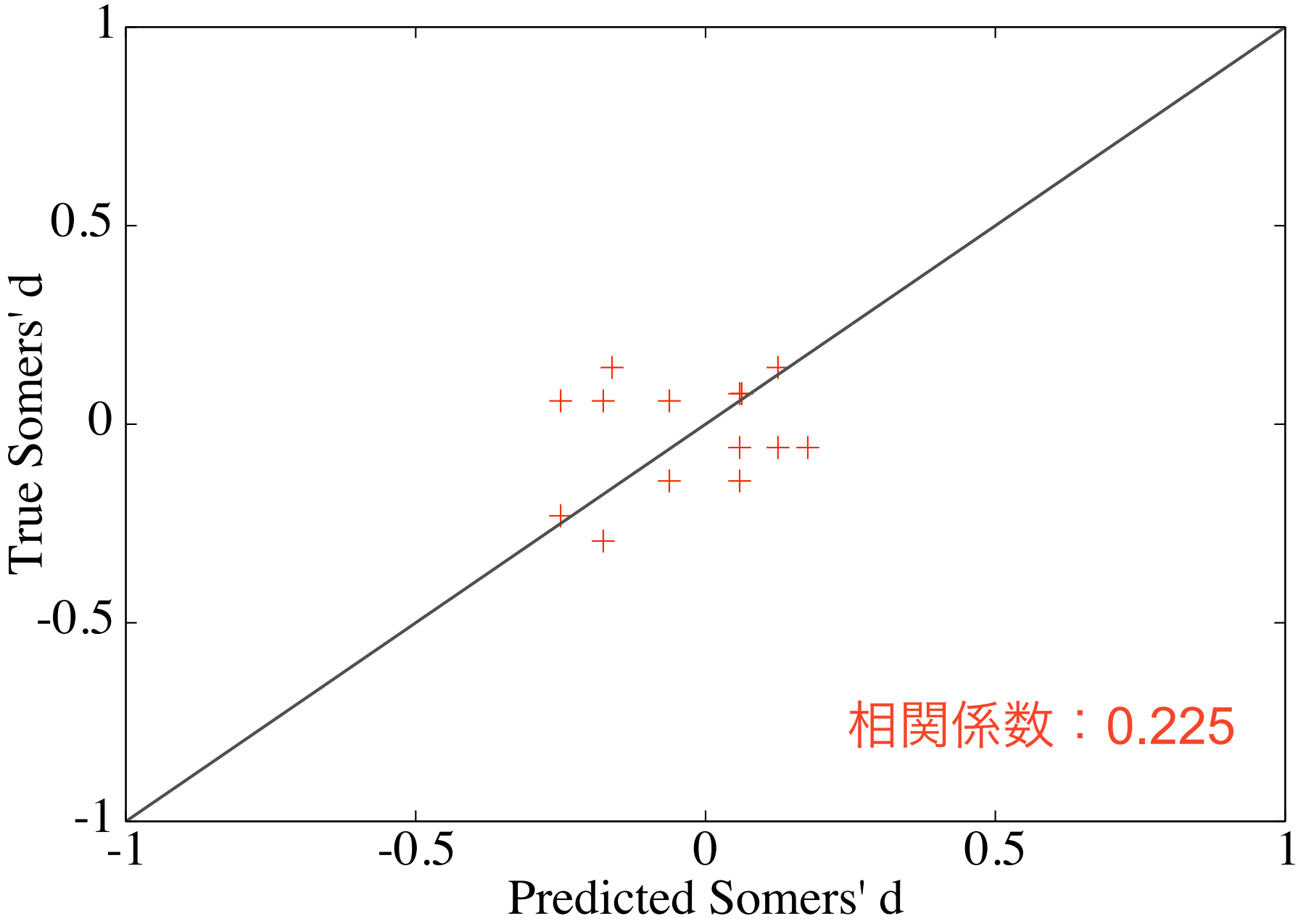
結果1：擬似ユーザによる推定区間の変化



結果2：擬似ユーザ数 k の変化による区間の変化



結果3： 真の類似度と推定した類似度の分布



まとめ

- ３組織間，垂直分割の環境における，ユーザ間類似度の計算方法を提案

- 通信コストをn^2からknに改善
- 複数のハブユーザを使い推定区間を集約