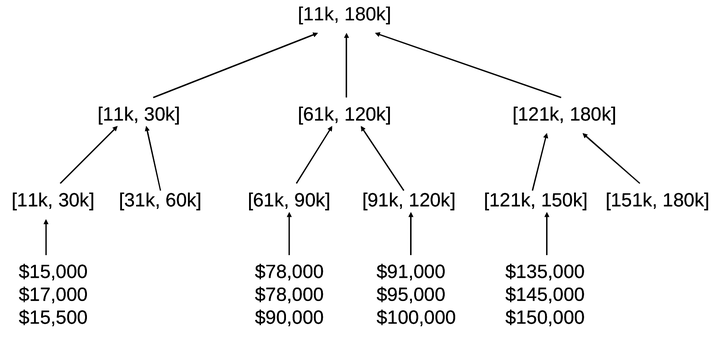
CERIAS

Privacy-Preserving Distributed *k*-Anonymity

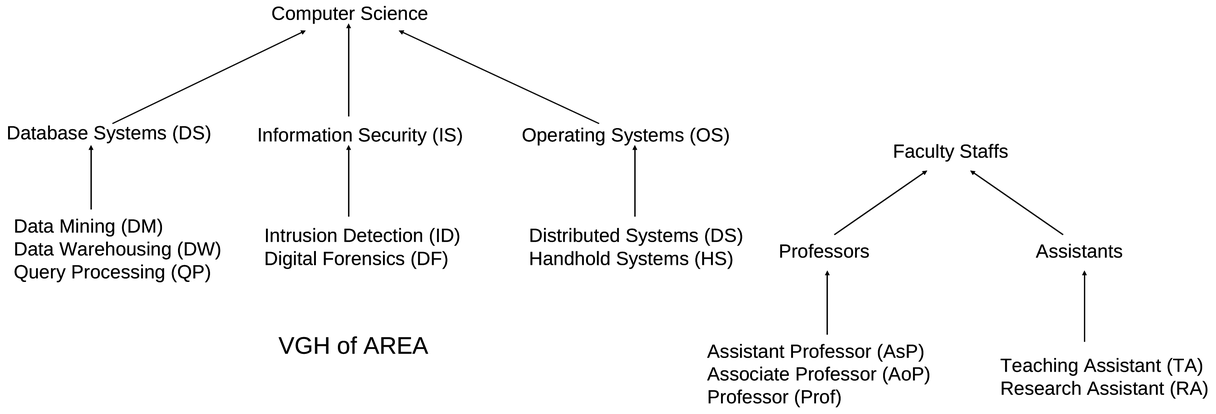
Wei Jiang & Chris Clifton

- What is *k*-Anonymity?
  - Let $T(A_1, \ldots, A_n)$ be a table and QI_T be the quasi-identifier set. T satisfies *k*-anonymity if and only if each sequence of values in $T[QI_T]$ appears at least *k* times.
- Why do we need *k*-Anonymity?
  - Protecting privacy: Limiting the ability to link released data to other external information
  - Preserving true data values and correlations among the original data
- Why distributed *k*-Anonymity?
  - Large amount of data are often stored at different sites
  - Privacy can be further enhanced
  - Different sites may not share their data directly
- How to achieve *k*-Anonymity?
  - Suppression: A sensitive value can be hidden by a * (mask)
  - Generalization: A specific value can be substituted by a more general value according to some Value Generalization Hierarchies (VGHs), i.e. $135,000 can be generalized to interval [121k, 150k]
  - Existing systems: Datafly, *k*-Similar, *μ*-argus, etc
- What is our contribution?
  - Propose a *k*-anonymity protocol that produces global *k*-anonymous data when the dataset is vertically partitioned between two sites
  - The protocol preserves *k*-anonymity (or privacy) between the sites: It learns nothing that violates *k*-anonymity with respect to the data at the other site

Examples of Value Generalization Hierarchies (VGHs):



VGH of SALARY



VGH of AREA

VGH of POSITION

Protocol Execution: Round 1

ID	AREA	POSITION	SALARY	SSN
1	Data Mining	Associate Professor	$90,000	708-79-1698
2	Intrusion Detection	Assistant Professor	$91,000	606-67-6789
3	Data Warehousing	Associate Professor	$95,000	626-23-1459
4	Intrusion Detection	Assistant Professor	$78,000	373-55-7788
5	Digital Forensics	Professor	$150,000	626-87-6503
6	Distributed Systems	Research Assistant	$15,000	708-66-1552
7	Handhold Systems	Research Assistant	$17,000	810-74-1079
8	Handhold Systems	Research Assistant	$15,500	606-37-7706
9	Query Processing	Associate Professor	$100,000	373-79-1698
10	Digital Forensics	Assistant Professor	$78,000	999-03-7892
11	Digital Forensics	Professor	$135,000	708-90-1976
12	Intrusion Detection	Professor	$145,000	606-17-6512

Original Dataset before Partitioning

ID	$AREA^1$	$POSITION^0$
1	Database Systems	Associate Professor
2	Information Security	Assistant Professor
3	Database Systems	Associate Professor
4	Information Security	Assistant Professor
5	Information Security	Professor
6	Operating Systems	Research Assistant
7	Operating Systems	Research Assistant
8	Operating Systems	Research Assistant
9	Database Systems	Associate Professor
10	Information Security	Assistant Professor
11	Information Security	Professor
12	Information Security	Professor

P1's locally *k*-anonymous data

$y_1 = \{ \{1,3,9\}, \{2,4,10\}, \{5,11,12\}, \{6,7,8\} \}$

$y_2 = \{ \{1,4,10\}, \{2,3,9\}, \{5,11,12\}, \{6,7,8\} \}$

$y_1 \neq y_2$ since $|\{1,3,9\} \cap \{2,3,9\}| = 2 < k = 3$

ID	$SALARY^1$
1	[61k, 90k]
2	[91k, 120k]
3	[91k, 120k]
4	[61k, 90k]
5	[121k, 150k]
6	[11k, 30k]
7	[11k, 30k]
8	[11k, 30k]
9	[91k, 120k]
10	[61k, 90k]
11	[121k, 150k]
12	[121k, 150k]

P2's locally *k*-anonymous data

Algorithm: DPP_2GA

Require: Private Data T1, QI = {A1,…,An}, Constraint *k*, Hierarchies VGH_{Ai}, where i = 1, …, n, assume $k \leq |T1|$

1. P1 generalizes his date to be locally k-anonymous
2. int c = 0;
3. **repeat**
4. c = c + 1**;**
5. P1 computes γ_c;
6. P1 computes $E_{K1}(\gamma_c)$ and sends it to P2;
7. P1 receives $E_{K2}(\gamma_c')$ and computes $\Gamma_{P2} = E_{K1}(E_{K2}(\gamma_c'))$;
8. P1 receives $\Gamma_{P1} = E_{K2}(E_{K1}(\gamma_c))$;
9. **until** $\Gamma_{P1} = \Gamma_{P2}$
10. return $T_k[QI] = T1_{\gamma c} \bowtie T2_{\gamma c}$.

Protocol Execution: Round 2

ID	$AREA^1$	$POSITION^1$
1	Database Systems	Professors
2	Information Security	Professors
3	Database Systems	Professors
4	Information Security	Professors
5	Information Security	Professors
6	Operating Systems	Assistant
7	Operating Systems	Assistant
8	Operating Systems	Assistant
9	Database Systems	Professors
10	Information Security	Professors
11	Information Security	Professors
12	Information Security	Professors

P1's locally *k*-anonymous data

$y_1 = \{ \{1,3,9\}, \{2,4,5,10,11,12\}, \{6,7,8\} \}$

$y_2 = \{ \{1,2,3,4,9,10\}, \{5,11,12\}, \{6,7,8\} \}$

$y_1 = y_2$

ID	$SALARY^2$
1	[61k, 120k]
2	[61k, 120k]
3	[61k, 120k]
4	[61k, 120k]
5	[121k, 180k]
6	[11k, 30k]
7	[11k, 30k]
8	[11k, 30k]
9	[61k, 120k]
10	[61k, 120k]
11	[121k, 180k]
12	[121k, 180k]

P2's locally *k*-anonymous data

k-Anonymous Dataset, where *k* = 3 and QI = {AREA, POSITION, SALARY}

Key Observations of DPP_2GA

- If $T_k[QI]$ is *k*-anonymous, then $T_k[QI']$ is also *k*-anonymous, where QI' □ QI
- T_k achieved through generalization satisfies *k*-anonymity if and only if □t′ □ T_k, Prob[t′ ▷ t □ T] ≤ 1/*k*
- If $y_1 = y_2$, then there are no *p*, *q* such that $0 < |y_1[p] \cap y_2[q]| < k$
- If $y_1 = y_2$, then $T_k[QI] = T1_{y1} \triangleright\triangleleft T2_{y2}$ satisfies global *k*-anonymity
- Inference problems do exist when the equality test is performed between y_1 and y_2 (or Γ_{P1} and Γ_{P2}) at step 9 of the algorithm
- We have proved that the inference problems do not make the final resulting data less *k*-anonymous

PURDUE UNIVERSITY

CERIAS®

Discovery Park e-Enterprise Center