Consiglio Nazionale delle Ricerche

PURDUE UNIVERSITY

Hiding the Presence of Individuals from Shared Databases: δ-Presence

M. Ercan Nergiz

Maurizio Atzori

Chris Clifton

Pisa KDD Lab

Indiana Center for Database Systems

Outline

- Adversary Models
  - Existential Uncertainty Model
- δ-Presence
  - Checking for δ-Presence Property
  - Providing δ-Presence
- Future Work

Adversary Models

Original Dataset

Age	Sex	Address	Disease
17	M	W. Lafayette	Obesity
16	M	Lafayette	Obesity
23	F	Lafayette	Tetanus
25	F	Indianapolis	Flu

Adversary:
"I know that Chris is 'Male', from 'W. Lafayette' and 17-year-old.
What is his disease?"

k-Anonymity

Age	Sex	Address	Disease
15-18	M	G. Lafayette	Obesity
15-18	M	G. Lafayette	Obesity
22-26	F	Indiana	Tetanus
22-26	F	Indiana	Flu

"Chris is definitely obese."

Adversary Models

l-Diversity, t-Closeness

Age	Sex	Address	Disease
15-26	*	Indiana	Obesity
15-26	*	Lafayette	Obesity
15-26	*	Lafayette	Tetanus
15-26	*	Indiana	Flu

Adversary:
"Chris is not necessarily obese."

Anatomization

Age	Sex	Address	Disease
17	M	W. Lafayette	{Ob,Flu}
16	M	Lafayette	{Ob,Te}
23	F	Lafayette	{Ob,Te}
25	F	Indianapolis	{Ob,Flu}

Adversary:
"Chris is *still* not necessarily obese."

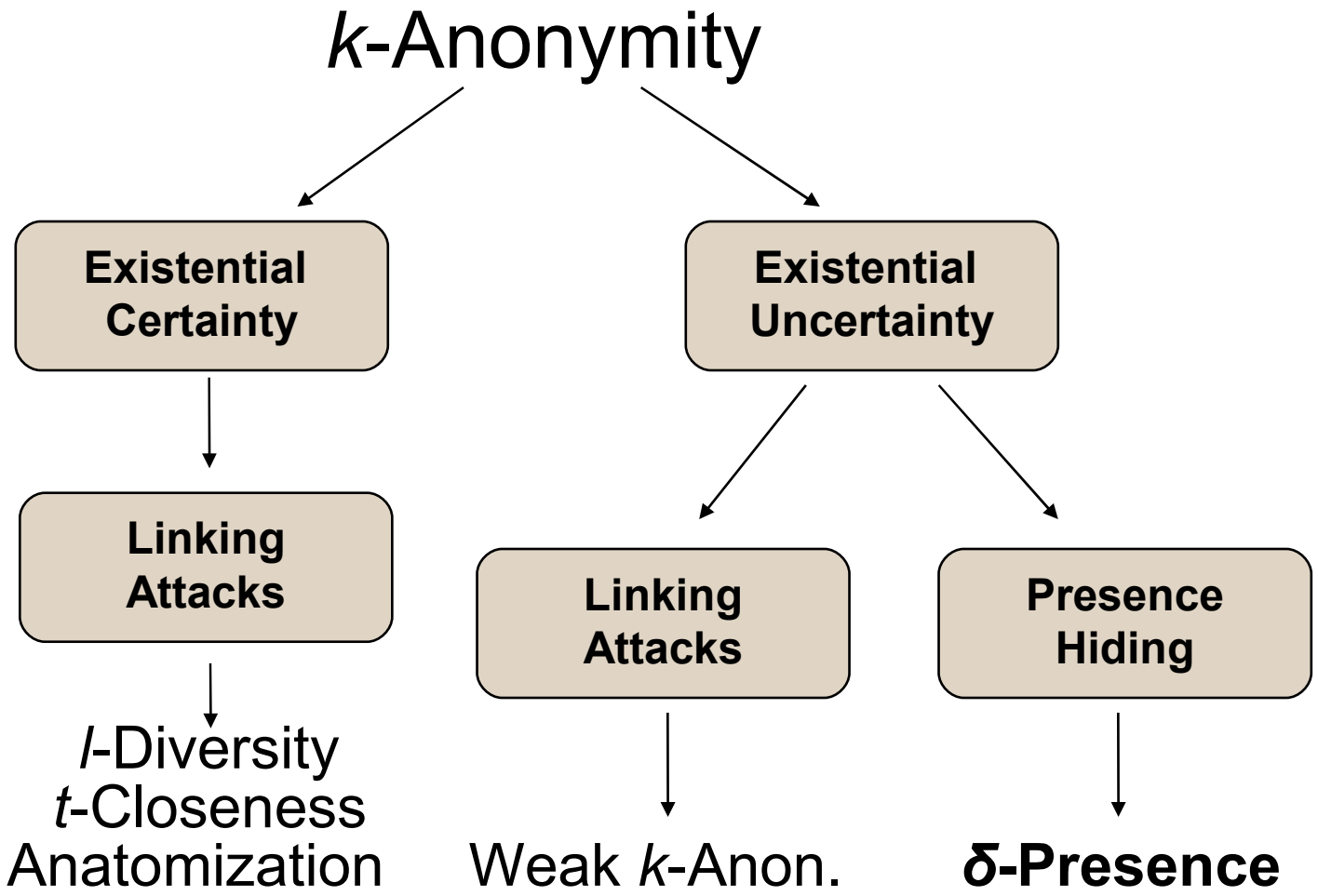
Adversary Models and Possible Threats

- **Existential Certainty:** Adversary knows that the individual is in the private dataset and tries to learn the sensitive information about the individual in the private dataset.
  - **Linking Attacks:** Linking Identities with sensitive attributes
- **Existential Uncertainty:** Adversary doesn't know the individual is or is not in the private dataset.
  - **Linking Attacks:** Existential disclosure is not considered as a privacy violation given that sensitive information is protected according to given privacy constraints.
  - **Presence Hiding:** Disclosure of existence or absence of an individual in the private dataset is a privacy violation.

k-Anonymity

- Provides *some* protections for all of the adversary models.
  - Sensitive info protection
  - Identity protection by QI anonymizations
- **BUT** is not perfect for any of the models

k-Anonymity Extensions



δ-Presence

- The risk is simply from identifying that an individual is (or is not) in an anonymized dataset.
- Can be interpreted in terms of increased risk of disclosure.
- A meaningful bridge between human-understandable policy and mathematically sound standards for anonymity.
  - E.g., can we speak of privacy in terms of risk/cost/benefit?
  - Can convert $ to *δ (see paper).*

δ-Presence

Given an external (public) background knowledge P, and a private table T;

$\delta = (\delta_{min}, \delta_{max})$-presence holds

for a generalization T^* of T if

$$\delta_{min} \leq Pr(t \in T \mid T^*, P) \leq \delta_{max}$$

for every $t \in P$

Presence Challenge

P

	Publicly Known Data				
	Name	Zip	Age	Nationality	*Sen.*
a	Alice	47906	35	USA	*0*
b	Bob	47903	59	Canada	*1*
c	Christine	47906	42	USA	*1*
d	Dirk	47630	18	Brazil	*0*
e	Eunice	47630	22	Brazil	*0*
f	Frank	47633	63	Peru	*1*
g	Gail	48973	33	Spain	*0*
h	Harry	48972	47	Bulgaria	*1*
i	Iris	48970	52	France	*1*

T

	Research Subset		
	Zip	Age	Nationality
b	47903	59	Canada
c	47906	42	USA
f	47633	63	Peru
h	48972	47	Bulgaria
i	48970	52	France

How to find δ-present generalization of *T*?

Checking for Presence Property: Non-overlapping Generalization

- A generalization T^* of T is a non-overlapping generalization w.r.t. P if
  - every tuple in P can be mapped onto at most one equivalence class in T^* .
- Checking presence property for non-overlapping generalizations is easy

Checking for Presence Property: Non-overlapping Generalization Ex.

P

	Publicly Known Data				
	Name	Zip	Age	Nationality	*Sen.*
a	Alice	47906	35	USA	*0*
b	Bob	47903	59	Canada	*1*
c	Christine	47906	42	USA	*1*
d	Dirk	47630	18	Brazil	*0*
e	Eunice	47630	22	Brazil	*0*
f	Frank	47633	63	Peru	*1*
g	Gail	48973	33	Spain	*0*
h	Harry	48972	47	Bulgaria	*1*
i	Iris	48970	52	France	*1*

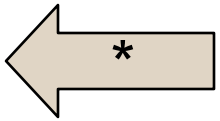
T*

	Research Subset		
	Zip	Age	Nationality
b	47*	*	America
c	47*	*	America
f	47*	*	America
h	48*	*	Europe
i	48*	*	Europe

Checking for Presence Property: Non-overlapping Generalization Ex.

P*

	Public Dataset			Sen.
	Zip	Age	Nationality	
a	47*	*	America	0
b	47*	*	America	1
c	47*	*	America	1
d	47*	*	America	0
e	47*	*	America	0
f	47*	*	America	1
g	48*	*	Europe	0
h	48*	*	Europe	1
i	48*	*	Europe	1



T*

	Research Subset		
	Zip	Age	Nationality
b	47*	*	America
c	47*	*	America
f	47*	*	America
h	48*	*	Europe
i	48*	*	Europe

Checking for Presence Property

- Let T^* be a non-overlapping generalization of T w.r.t. P. Then T^* is δ-present, if for each equivalence class ec of the corresponding P^*:

$$\delta_{min} \leq (\text{\# of 1s in Sen.}) / |ec| \leq \delta_{max}$$

(.5-.66)-Presence

P*

	Public Dataset			Sen.
	Zip	Age	Nationality	
a	47*	*	America	0
b	47*	*	America	1
c	47*	*	America	1
d	47*	*	America	0
e	47*	*	America	0
f	47*	*	America	1
g	48*	*	Europe	0
h	48*	*	Europe	1
i	48*	*	Europe	1

T*

	Research Subset		
	Zip	Age	Nationality
b	47*	*	America
c	47*	*	America
f	47*	*	America
h	48*	*	Europe
i	48*	*	Europe

$Pr(t_a \in T \mid T^*) = 0.5$

$Pr(t_g \in T \mid T^*) = 0.66$

k-Anonymity Fails

P*

	Publicly Released Dataset			
	Zip	Age	Nationality	Sen.
a	4*	≤ 40	*	0
d	4*	≤ 40	*	0
e	4*	≤ 40	*	0
g	4*	≤ 40	*	0
b	4*	> 40	*	1
c	4*	> 40	*	1
f	4*	> 40	*	1
h	4*	> 40	*	1
i	4*	> 40	*	1

5-anonymous T*

	Research Subset		
	Zip	Age	Nationality
b	4*	> 40	*
c	4*	> 40	*
f	4*	> 40	*
h	4*	> 40	*
i	4*	> 40	*

$Pr(t_a \in T \mid T^*) = 0$

$Pr(t_b \in T \mid T^*) = 1$

How to Provide Presence?: Anti-monotonicity

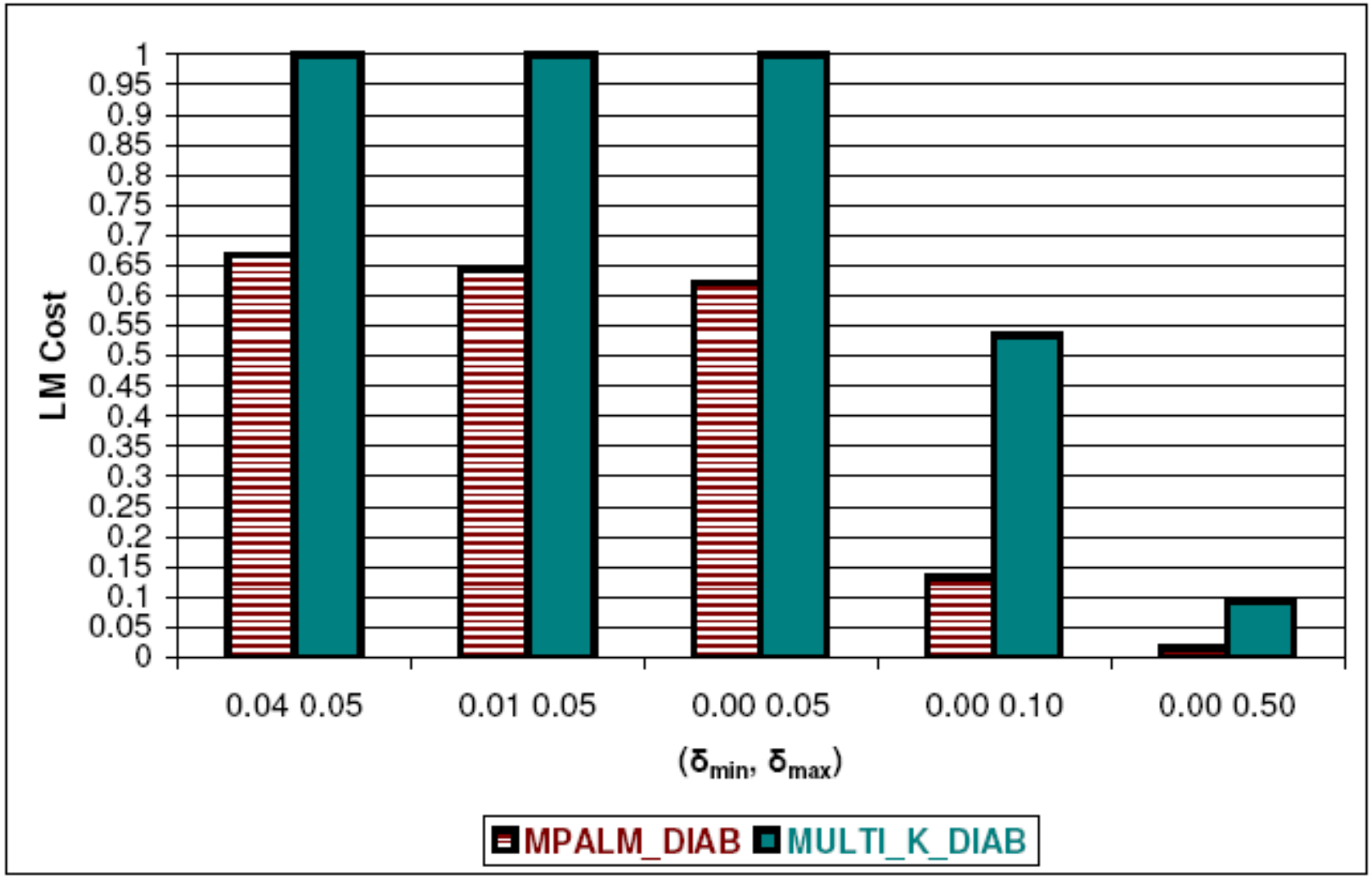
- Given a public table P, private table T, a non-overlapping generalization T_1^* of T, and a non-overlapping generalization T_2^* of T_1^*.

  If T_2^* is not δ-present w.r.t. P and T then neither is T_1^*.

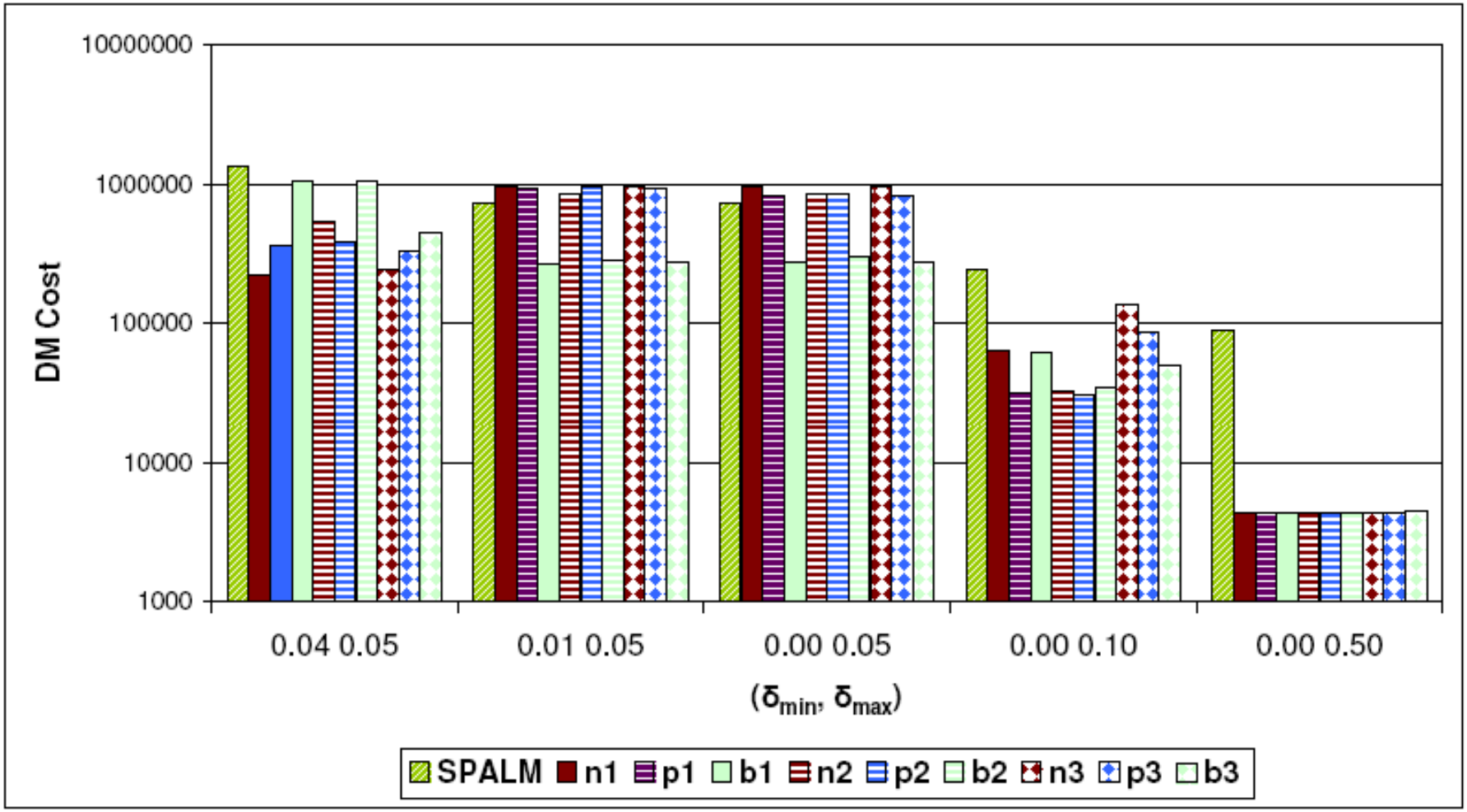
How to Provide Presence?: SPALM, MPALM

- SPALM: Optimum Single Dim. Presence Alg.
  - Analogous to Incognito *[LDR SIGMOD05]*
  - Top down pruning approach
- MPALM: Multi Dim. Presence Alg.
  - Analogous to Mondrian *[LDR ICDE06]*
  - With different attribute selection heuristics

Experiments



Experiments



Future Work

- Assume distribution of attributes instead of a public table.
- Apply randomization on private table *T* to satisfy presence.
- Design a clustering based presence algorithm with overlapping equivalence classes.
- Assume sensitive attributes exist in *T*
- Make risk analysis on the selection of δ parameters w.r.t. real world scenarios.
- Personalize privacy based on attributes of the individuals.

Hiding the Presence of Individuals from Shared Databases: δ-Presence

Thanks for listening

atzori@di.unipi.it

Questions?