ADVANCED DATA MANAGEMENT

Gianluca Oldani – Tutoring 4

What will be discussed today

- Refresh K-Anonymity + L-Diversity
- The Mondrian algorithm (theory + Spark impl)
- Distributed Job Scheduling

What is K-Anonymity

- Problem statement:
  Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful

What is K-Anonymity

- Property to enforce:
  the information for each person contained in the release cannot be distinguished from at least **k- 1** individuals whose information also appear in the release

What is K-Anonymity

Is this dataset k-anonymous for k = 2?

SSN	Name	Race	Date of birth	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

What is K-Anonymity

The previous question is not meaningful, since it is not specified what parts of the table has to be 2-anonymous

SSN	Name	Race	Date of birth	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

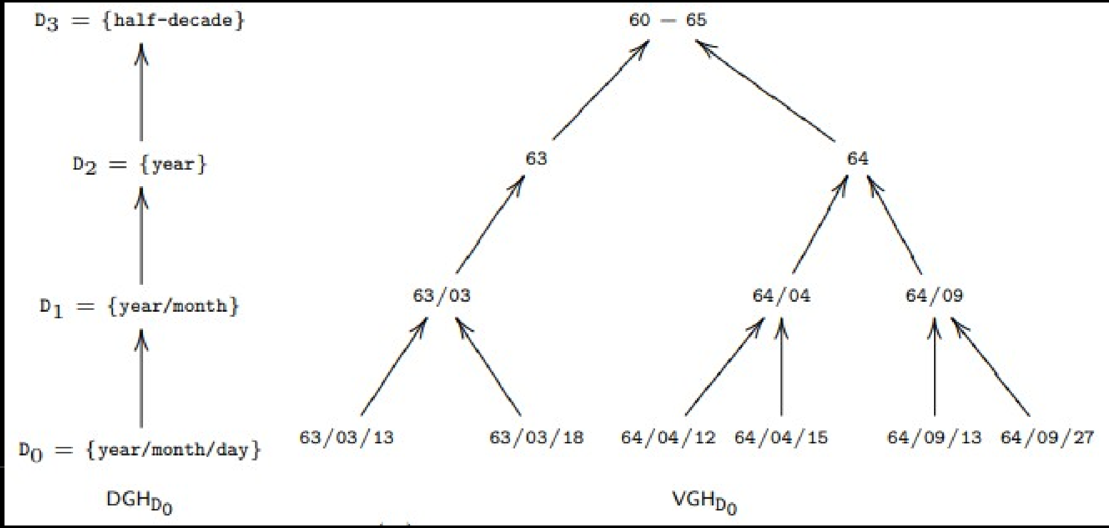
The attributes that can be used by an attacker to performa a linkage attack are calles **quasi-identifiers**

Why such a model is necessary?

- Popular concerns:
  - Netflix published data about movie rankings for 500,000 customers in 2007, and researchers showed they could de-anonymize the data using a few additional inputs from IMDb
  - Using 1990 U.S. census data, Stanford researchers showed that they could uniquely identify **87 percent** of the U.S. population using only their Zip code, gender, and date of birth
  - AOL published search data for 650,000 users in 2007, thinking it was enough to anonymize their name using a unique ID. Unfortunately, most users often query their own name. As a result, their CTO resigned and an entire research team was fired after the public outcry

How to achieve K-Anonymity

- Suppression: remove the tuple from the dataset
- Generalization:substituting the values of a given attribute with more general values



Introducing L-Diversity

- Some attributes can be valuable as much as identifiers
- K-Anonymity does not enforce any constraint on the cardinality of the set of values in a group of K individuals

Introducing L-Diversity

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

The Mondrian algorithm

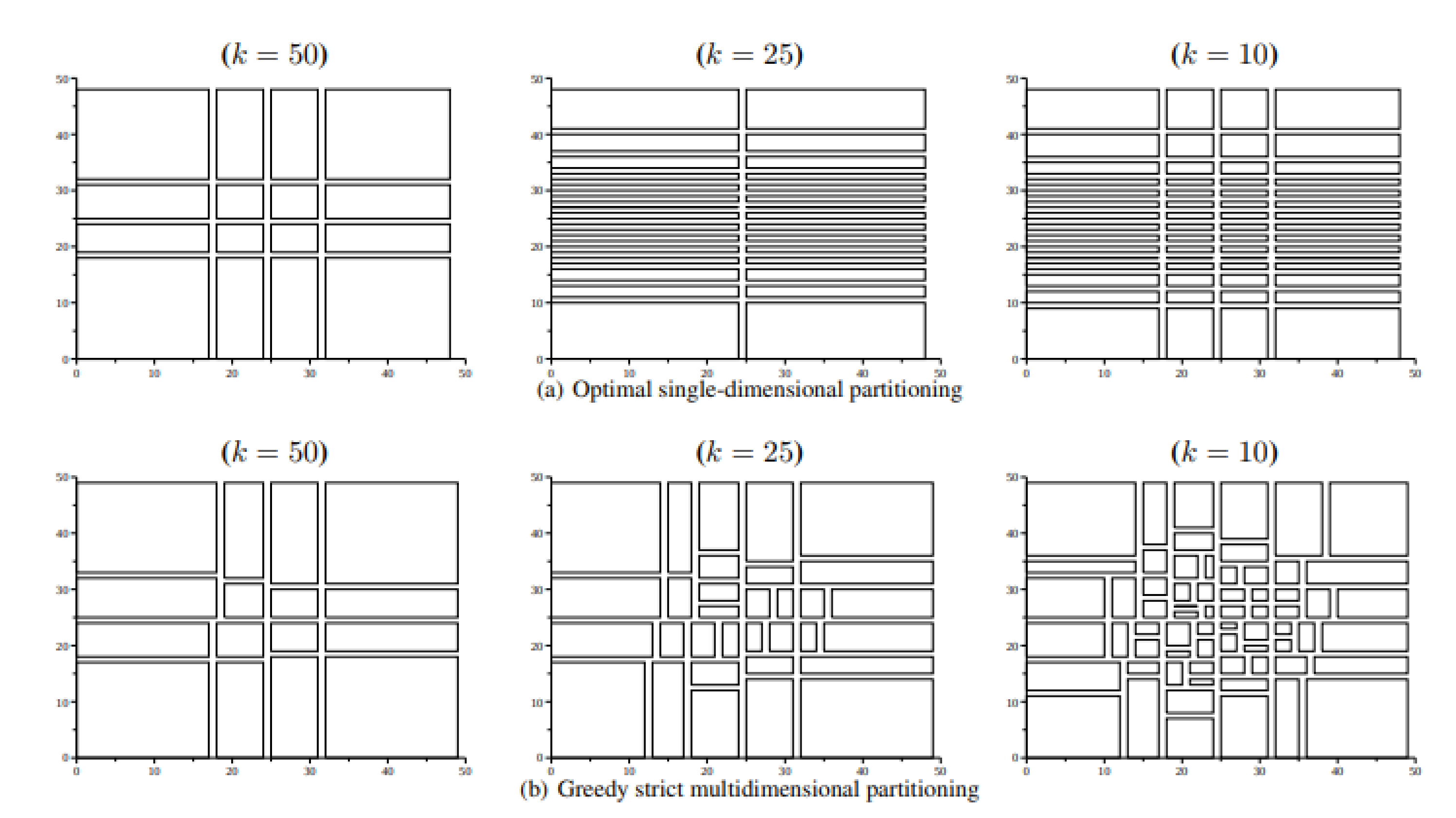
- Greedy approach to enforce K-Anonymity + L-Diversity
- More efficient than the optimal single attribute approach (nLog(n) vs exp(n))
- Reported to produce better results in the multidimensional scenario

The Mondrian algorithm

Anonymize($partition$)
  if (no allowable multidimensional cut for $partition$)
    return $\phi : partition \rightarrow summary$
  else
    $dim \leftarrow$ choose_dimension()
    $fs \leftarrow$ frequency_set($partition$, dim)
    $splitVal \leftarrow$ find_median(fs)
    $lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$
    $rhs \leftarrow \{t \in partition : t.dim > splitVal\}$
    return Anonymize(rhs) $\cup$ Anonymize(lhs)

- Strict version: lhs and rhs have no values in common

Lesson learned: visualization is important



Mondrian components

- choose_dimension() → normalized span
- frequency_set() → scan
- find_median() → sort and filter
- hidden => check partition validity → scan

Reality check – Problems of Mondrian

- choose_dimension() → O(M * N)
- frequency_set() → O(N)
- find_median() → O(FLog(F))
- hidden => check partition validity → O(1) for K-Anonymity, O(K*N) for L-Diversity

- N = tuples in the partition
- M = # of quasi-identifiers
- F = # unique values
- K = # of sensitive attributes

Reality check – Problems of Mondrian

- choose_dimension() → A lot of quasi-identifiers?
- frequency_set() → Large datasets?
- find_median() → F → N ?

- What about attributes without order?

Spark to the rescue

DEMO TIME!

Repository link: https://github.com/mosaicrown/mondrian

Papers links:

https://spdp.di.unimi.it/papers/percom2021.pdf

https://spdp.di.unimi.it/papers/percom2021-artifact.pdf

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9894678

Job Scheduling – Final push for today

- Problem of normal job scheduling?
- Problem of job scheduling over a cluster?
- Any idea for the architecture to use?