Information Integration – Exercise 7 – Gabriel Glaser

Task 3: Entity Resolution and Fusion (Sample exam question)

a) For relational entity resolution, briefly describe the naive algorithm. What is its complexity?

b) What algorithm to reduce complexity did we discuss? What is the runtime complexity of this approach for typical entity resolution tasks?

c) Which conflict types for data fusion have we introduced in class? Provide a brief example for each of them.

d) We have discussed in detail how to implement **complementation**. Given the source data shown in the table below, apply the algorithm to obtain all maximal complementing sets. Your answer should include the **tree structure** constructed by the algorithm with marked **maximal complementing sets**. Note that the first attribute of the table are tuple identifiers that you can use to refer to the tuples.

tid	A	B	C	D	E
t_1	M	1	x	A	$\perp$
t_2	M	1	$\perp$	A	$\perp$
t_3	P	2	$\perp$	B	$\perp$
t_4	P	2	$\perp$	$\perp$	2
t_5	P	2	x	B	2
t_6	M	$\perp$	x	$\perp$	1
t_7	M	1	x	$\perp$	$\perp$
t_8	P	$\perp$	x	C	2

a) Pair-wise comparison of each entity (e.g., similarity measure + threshold) based on Cartesian product. This has a quadratic runtime complexity.

b) Sorted-Neighbourhood reduces the complexity to $\mathcal{O}(n \log n)$. Calculate n keys, sort $(n \log n)$ and linear scan to find duplicate candidates in a constant window.

c)

- *Exact duplicate*: No problem, can drop one of the entities (SQL UNION). For example,

title	author	year
Harry Potter 1	J.K. Rowling	1996
Harry Potter 1	J.K. Rowling	1996

.

- *Subsumption*: One entity is the subset of the other entity. For example,

title	author	year
Harry Potter 1	J.K. Rowling	1996
Harry Potter 1	J.K. Rowling	

.

- *Complementation*: Either both entities contain the same value or only one entity has a value. For example,

title	author	year
Harry Potter 1		1996
Harry Potter 1	J.K. Rowling	

.

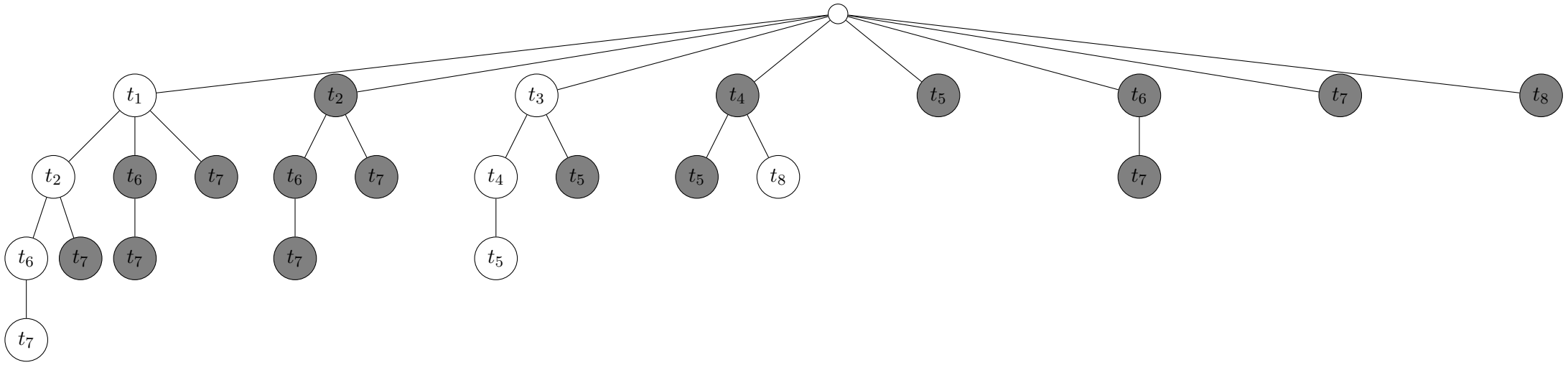
- D*ata Conflict*: There is an attribute where two different values are provided. For example,

title	author	year
Harry Potter 1	J.K. Rowling	1996
Harry Potter and the Sorcerer's Stone	J.K. Rowling	

.

d)

- t_1, t_2, t_6, t_7
- t_3, t_4, t_5
- t_4, t_8


t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_2
t_6
t_7
t_6
t_7
t_4
t_5
t_5
t_8
t_7
t_6
t_7
t_7
t_7
t_5
t_7