Data Engineering Exercise Sheet 04

1. Big Data Integration: Overview

Big Data Integration combines Big Data with data integration. This task provides an overview over the challenges of Big Data Integration.

(a) Before we discuss Big Data Integration, we have a look at data integration. What are the three major steps of data integration that were introduced in class? Describe each of the three steps.

1. Schema Alignment: Mapping of the source schemas into one global schema. (Integrated schema)

2. Record Linkage: Mapping the data to an entity and finding duplicates.

3. Data Fusion: Merge duplicates and format data with defined rules.

(b) We are given two tables about musicians. Employ the three steps of the data integration process to merge them into a single table. Describe each of your actions. Provide intermediate results (for each step).

Musicians 1					
First Name	Last Name	Born	Died	Height	Music Pieces
Joseph	Haydn	1732	1809	169	154
Schubert	Franz	1797	1828	183	604
Johann Sebastian	Bach	1685	1750	165	1128
Wolfgang	Mozart	1756	1791	175	626

Musicians 2				
Name	Date of Birth	Date of Death	Height	Pieces
Joseph Haydn	03/31/1732	05/31/1809	5'7"	207
Wolfgang Amadeus Mozart	01/27/1756	12/05/1791	5'9"	635
Franz Schubert	01/31/1797	11/19/1828	6'0"	598

1. Schema Alignment: Global Schema

First Name	Last Name	Birth	Death	Height	Music Pieces

2. Record Linkage
   Rules: First Name and Last Name must be the same + Birth Year and Death Year

First Name	Last Name	Birth	Death	Height	Music Pieces
Joseph	Hadyn	1732	1809	169	154
Joseph	Haydn	03/31/1732	05/31/1809	5’7”	207
Franz	Schubert	1797	1828	183	604
Franz	Schubert	01/31/1797	11/19/1828	6’0”	598
Johann Sebastian	Bach	1685	1750	165	1128
Wolfgang	Mozart	1756	1795	175	626
Wolfgang Amadeus	Mozart	01/27/1756	12/05/1791	5’9”	635

3. Data Fusion

Rules:

- First Name, Last Name, Birth and Death entries with the most information
- Height centimeters
- Music Pieces average

First Name	Last Name	Birth	Death	Height	Music Pieces
Joseph	Haydn	03/31/1732	05/31/1809	169	180
Franz	Schubert	01/31/1797	11/19/1828	183	601
Johann Sebastian	Bach	01/01/1685	31/12/1750	165	1128
Wolfgang Amadeus	Mozart	01/27/1756	12/05/1791	175	630

Take the higher value, because might be more updated, average also ok

Dates use default, uncertain when actually using the data

(c) Big Data imposes additional challenges on data integration. In class, three challenges are introduced. Describe each of these challenges.

Volume: Number of structures sources. Difficult to do exact schema alignment semi automatically and expensive.

Variety: Differences between sources regarding structure and completeness of data

Veracity: Consistency and Accuracy of Data between different sources

2. Big Data Integration: Probabilistic Mediated Schema Generation

A key challenge of Big Data Integration is schema alignment. Its goal is to automatically integrate multiple schemas from different Web sources. One approach to achieve this goal is probabilistic schema alignment. In this exercise we will discuss its first step, which is the automatic creation of a probabilistic mediated schema from a set of sources.

(a) In order to apply probabilistic schema matching to multiple datasets from the Web, some assumptions must be fulfilled. Describe the assumptions introduced in class.

Schema of sources are given (not always given)

We deal with relation (flat) schemas (widely spread, but not all sources have it)

We aim for a domain-independent solution

(b) Let us assume we have to align the schemas of six sources containing contact information $\{S_1,...,S_6\}$. Their schemas are listed in the following table: Define the set of all (distinct) source attributes $A = \{a_1,...,a_m\}$ from the above schemas $\{S_1,...,S_6\}$.

Schema	Layout
S_1	name, mobile, phone, address, email address
S_2	name, mobile, private address, email address
S_3	name, mobile phone, office phone, address, home address, email
S_4	name, mobile phone, phone, email
S_5	name, mobile, secondary phone, private address
S_6	customer-id, mobile, home address, email address

Set of distinct attributes = {name, mobile, phone, address, email address, private address, mobile phone, office phone, home address, email, secondary phone, customer-id}

(c) Let us find out which attributes occur frequently in the sources. Calculate the frequency for each attribute a_j. In our example, the frequency f is defined as:

$$f(a_j) = \frac{|\{i \in [1,6] | a_j \in S_i\}|}{6}$$

$f(a_j)$	frequency
name	5/6
mobile	4/6
phone	1/6
address	2/6
email address	3/6
private address	2/6
mobile phone	2/6
office phone	1/6
home address	2/6
email	2/6
secondary phone	1/6
customer-id	1/6

(d) Next, we remove infrequent attributes from *A*. By our definition, infrequent items have a frequency smaller than Θ = 1/3 . Write down the reduced content of *A*.

$f(a_j)$	frequency
name	5/6
mobile	4/6
address	2/6
email address	3/6
private address	2/6
mobile phone	2/6
home address	2/6
email	2/6

(e) Let us assume we have calculated the following similarity measures *s* for any two attributes a_j and a_k from *A*:

$s(a_j, a_k)$	name	mobile	mobile phone	phone	address	private address	home address	email address	email
name	1	0,11	0,13	0,09	0,32	0,27	0,25	0,65	0,72
mobile		1	0,88	0,66	0,22	0,11	0,21	0,34	0,32
mobile phone			1	0,77	0,20	0,15	0,16	0,29	0,23
phone				1	0,33	0,19	0,14	0,19	0,22
address					1	0,92	0,67	0,43	0,03
private address						1	0,90	0,33	0,40
home address							1	0,29	0,35
email address								1	0,87
email									1

We consider two attributes a_j and a_k similar if their similarity measure $s(a_j,a_k)$ is bigger or equal to threshold τ. Since we like to refine our result in the following step we allow for an error . We set $\tau = 0.7$ and $\epsilon = 0.1$. Find and mark all similarity measures in the above table, for which the equation $s(aj, ak) \geq \tau - \epsilon$ holds.

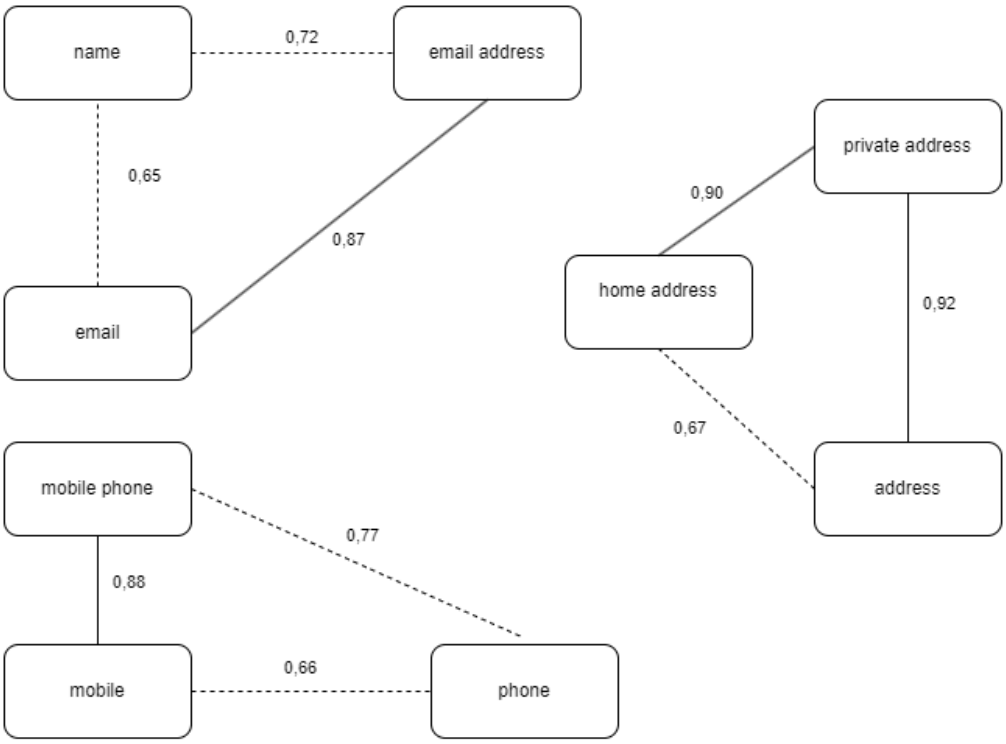
(f) Draw a graph $G(V,E)$ with

$$V = A$$

and

$$E = \{(aj, ak) | aj, ak \in A \wedge j/ = k \wedge s(aj, ak) \geq \tau - \epsilon\}$$

Intuitively the vertices of the graph are the attributes. The edges of the graphs are defined by the previously marked similarity measures.



(Edge: Email address und Email vertauscht)

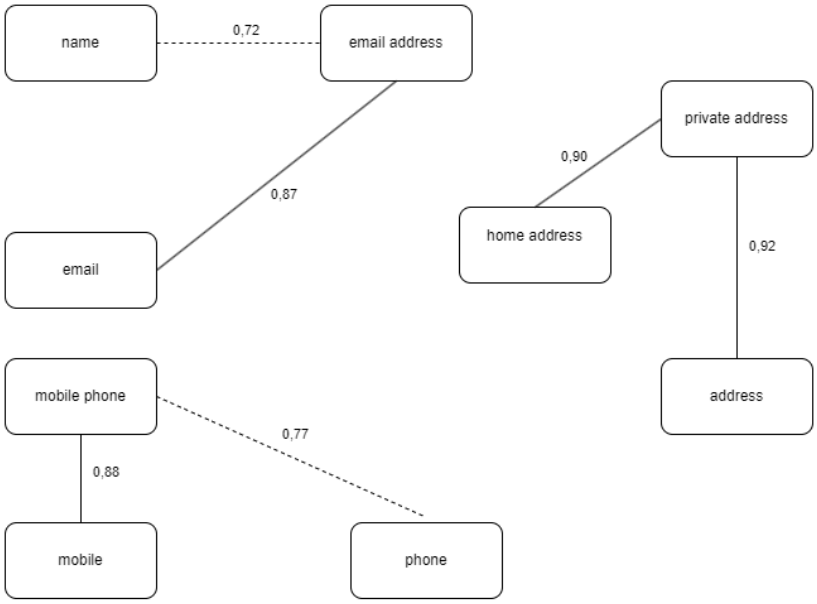
(g) The next step is to identify and mark all so called uncertain edges in the table and the graph. Uncertain edges are those edges whose similarity measure *s* is smaller than $\tau + \epsilon$, formally $\tau + \epsilon \geq s(a_j, a_k) \geq \tau - \epsilon$

Figure 1: Remove uncertain (dotted) edges when a path of certain (strong) edges exist

Figure 2: Remove uncertain (dotted) edges based on their similarity measure

See f) and h)

(h) It is time to remove some uncertain edges from the graph. First, remove all uncertain edges for which an alternative path with only certain edges exist (cf. Figure 1). Also, remove all uncertain edges for which a path of certain and at most one uncertain edge exist. Remove the uncertain edge with the lower similarity measure (cf. Figure 2).



(i) The next task is to create the alternative mediated schemas from the previously created graph. For example, the alternative mediated schemas from Figure 2 are:

$$M_1^* = \{\{a_1, a_2\}, a_3\}$$

$$M_2^* = \{\{a_1, a_2, a_3\}\}$$

$$M1 = (\{name\}, \{email\ address, email\}, \{private\ address, address, home\ address\}, \{mobile\ phone, mobile\}, \{phone\})$$

$$M2 = (\{name, email\ address, email\}, \{private\ address, address, home\ address\}, \{mobile\ phone, mobile, phone\})$$

$$M3 = (\{name\}, \{email\ address, email\}, \{private\ address, address, home\ address\}, \{mobile\ phone, mobile, phone\})$$

$$M4 = (\{name, email\ address, email\}, \{private\ address, address, home\ address\}, \{mobile\ phone, mobile\}, \{phone\})$$

(j) Let us compute the probabilistic mediated schema as the next step. For example, $M*1$ from the previous is consistent with source schema $S*1 = \{a1, a3\}$, whereas $M*2$ is not. For each alternative mediated schema $\{M_1,...M_l\}$ count the number of consistent source schemas $\{S_1,...,S_6\}$.

Consistent: M und S haben keine Widersprüche, z. B. S1 und M2 sind nicht konsistenz.

M1 = 6

M2 = 1 (S6)

M3 = 4 (S2, S3, S5, S6)

M4 = 1 (S6)

Werte sind falsch

Two attributes in a cluster are not in the same source

(k) Given the number of consistent source schemas for each alternative mediated schema $\{M_1,...M_l\}$, compute pp lthe probability for each schema $Pr(M_i)$. It is defined as

$$Pr(M_i) = \frac{c_i}{\sum_{j=1}^{l} c_j}$$

Which mediated schema has the highest probability?

Pr(M1) = 6 / 12 <-

Pr(M2) = 1 / 12

Pr(M3) = 4 / 12

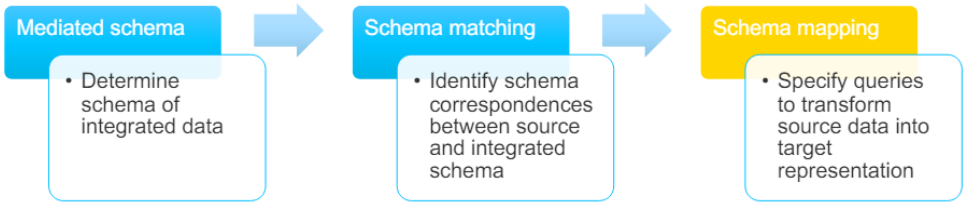
Pr(M4) = 1 / 12

3. Data Integration (Exam exercise)

a) When integrating data from multiple, heterogeneous sources, three major steps are required, namely schema alignment, record linkage, and data fusion. Briefly explain the goal of each step.

1. Schema Alignment: Mapping of the source schemas into one global schema.
2. Record Linkage: Mapping the data to an entity and finding duplicates.
3. Data Fusion: Merge duplicates and format data with defined rules.

b) Ignoring the issues of Big Data, what are the three essential steps of schema alignment? Name them and provide the goal of each of these steps.



c) To perform schema alignment on Big Data, we have discussed probabilistic schema alignment. Why are probabilities introduced? Which Big Data challenge(s) does the introduced probabilities address?

Volume: It is necessary for (semi)-automatic schema alignment, because there are uncertainties about semantics of attributes in the sources.

Variety: need to find the data that is most likely to be useful

Veracity: need to find the data that is most likely to be correct

d) An excerpt of the algorithm for probabilistic schema alignment discussed in class is shown on the next page. Concerning this algorithm, answer the following questions (using the space on the next page). Assumptions necessary to answer these are also provided on the next page.

(i) Complete the missing Step 7.

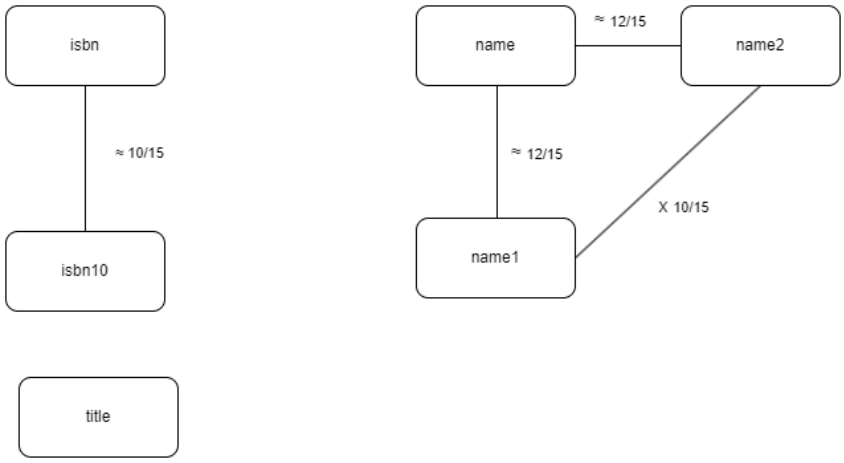
for each (subset of uncertain edges):

Omit the edges in the subset and compute a mediated schema where each connected component in the graph corresponds to an attribute in the schema;

(ii) Draw the graph resulting from Step 4 below. Make sure to include nodes, edges, and edge labels. Use solid lines for all edges.

AND (iii) In the graph you drew above, mark uncertain edges as determined in Step 5 using the symbol ≈.

AND (iv) In the graph you drew above, mark all edges removed in Step 6 using the symbol ×.



Name -> name1 certain edge

Name -> name2 certain edge

(v) Which mediated schemas are returned?

$M1 = \{\{isbn\}, \{isbn10\}, \{name\}, \{name1\}, \{name2\}, \{title\}\}$

$M2 = \{\{isbn, isbn10\}, \{name\}, \{name1\}, \{name2\}, \{title\}\}$

$M3 = \{\{isbn, isbn10\}, \{name, name1\}, \{name2\}, \{title\}\}$

$M4 = \{\{isbn, isbn10\}, \{name, name2\}, \{name1\}, \{title\}\}$

$M5 = \{\{isbn\}, \{isbn10\}, \{name, name1\}, \{name2\}, \{title\}\}$

$M6 = \{\{isbn\}, \{isbn10\}, \{name, name2\}, \{name1\}, \{title\}\}$

$M7 = \{\{isbn\}, \{isbn10\}, \{name, name1, name2\}, \{title\}\}$

$M8 = \{\{isbn, isbn10\}, \{name, name1, name1\}, \{title\}\}$

Nur 2, ISBN seperat und zusammen

(vi) Is this the result one would expect, given general knowledge about the book domain? If not, describe a schema that one may expect. How could you improve the algorithm to come closer to this ideal solution?

For this domain it works, e.g. there are several isbns (isbn-10, isbn-13) and authors (name1, name2)

0: **Input**: Source schemas $S_1, \ldots, S_n$.
Output: A set of possible mediated schemas.
1: Compute $\mathcal{A} = \{a_1, \ldots, a_m\}$, the set of all source attributes;
2: **for each** $(j \in [1, m])$
Compute frequency $f(a_j) = \frac{|\{i \in [1,n] | a_j \in S_i\}|}{n}$;
3: Set $\mathcal{A} = \{a_j | j \in [1, m], f(a_j) \geq \theta\}$; *//θ is a threshold*
4: Construct a weighted graph $G(V, E)$, where (1) $V = \mathcal{A}$, and (2) for each $a_j, a_k \in \mathcal{A}$, $s(a_j, a_k) \geq \tau - \epsilon$, there is an edge (a_j, a_k) with weight $s(a_j, a_k)$;
5: Mark all edges with weight less than $\tau + \epsilon$ as *uncertain*;
6: **for each** (uncertain edge $e = (a_1, a_2) \in E$)
Remove e from E if (1) a_1 and a_2 are connected by a path with only certain edges, or (2) there exists $a_3 \in V$, such that a_2 and a_3 are connected by a path with only certain edges and there is an uncertain edge (a_1, a_3);
7:
8: **return** distinct mediated schemas.

- We are given three source schemas:
$S1(isbn, title)$,
$S2(isbn10, name1, name2)$,
$S3(isbn, title, name)$.

- Set $\mathcal{A}$ produced in Step 3 retains all attributes.
- Threshold values are

$\tau = \frac{11}{15}$ and
$\epsilon = \frac{1}{15}$

- We are given two functions:
(1) $sharedPrefix(a_j, a_k, l)$ returns 1 if strings a_j and a_k share a prefix of length at least l and 0 otherwise
(2) $chars(a)$ returns the set of characters in string a,
Using these, we have

$$s(a_j, a_k) = sharedPrefix(a_j, a_k, 4) \times \frac{|chars(a_j) \cap chars(a_k)|}{|chars(a_j) \cup chars(a_k)|}$$

A = {isbn, title, isbn10, name1, name2, name}

$s(a_j, a_k)$	isbn	title	isbn10	name1	name2	name
isbn	1	0	4/6	0	0	0
title		1	0	0	0	0
isbn10			1	0	0	0
name1				1	4/6	4/5
name2					1	4/5
name						1