ulm university universität ulm

A Novel Approach on De-Identification of Heterogeneous Data based on a Modified Mondrian Algorithm

Master's Thesis

Fabian Singhofer | July 7, 2023 | Data Science and Big Data Analytics (DSBDA) in cooperation with BT (UK)

Outline

Motivation

Why do we want to anonymize data?

- Data and especially sharing it has become important nowadays
- **However:** Privacy regulations like the Data Protection Act 2018 in the UK require actions to enforce privacy
- **Particular action:** Anonymizing Personally Identifiable Information (PII)

What about heterogeneous data?

- Question came up during discussions with Aygul and Mathias
- Often there is a mix between traditional relational data and textual data

➔ **How can a combined anonymization approach look like?**

Motivating Example

A^*	Relational Attributes $A_1, ..., A_5$					Textual Attribute X
id	**gender**	**age**	**topic**	**sign**	**date**	**text**
1	male	36	Education	Aries	2004-05-14	My name is Pedro, I'm a 36 years old engineer from Mexico.
1	male	36	Education	Aries	2004-05-15	A quick follow up: I will post updates about my education in more detail.
2	male	24	Student	Leo	2005-08-18	I will start working for a big tech company as an engineer.
3	male	37	Banking	Pisces	2004-05-27	During my last business trip to Canada I met my friend Ben from college.
4	female	24	Science	Aries	2004-01-13	As a scientist from the UK, you can be proud!
4	female	24	Science	Aries	2004-01-17	Four days ago, I started my blog. Stay tuned for more content.
4	female	24	Science	Aries	2004-01-19	2004 will be a great year for science and for my career as a biologist.
5	male	29	indUnk	Pisces	2004-05-15	Did you know that Pisces is the last constellation of the zodiac.
6	female	27	Science	Aries	2004-05-15	Rainy weather again here in the UK. I hope you all have a good day!

- *RX*-dataset containing traditional relational attributes as well as free text attributes

A^* direct-identifying relational attribute
A_j quasi-identifying relational attribute
X textual attribute containing sensitive terms

Motivating Example

Structured Data

A^* | Relational Attributes $A_1, ..., A_5$

id	gender	age	topic	sign	date
1	male	36	Education	Aries	2004-05-14
1	male	36	Education	Aries	2004-05-15
2	male	24	Student	Leo	2005-08-18
3	male	37	Banking	Pisces	2004-05-27
4	female	24	Science	Aries	2004-01-13
4	female	24	Science	Aries	2004-01-17
4	female	24	Science	Aries	2004-01-19
5	male	29	indUnk	Pisces	2004-05-15
6	female	27	Science	Aries	2004-05-15

- Naive approach: Separation of concerns

A^* direct-identifying relational attribute
A_i quasi-identifying relational attribute
X textual attribute containing sensitive terms

Unstructured Data

A^* | Textual Attribute X

id	text
1	My name is Pedro, I'm a 36 years old engineer from Mexico.
1	A quick follow up: I will post updates about my education in more detail.
2	I will start working for a big tech company as an engineer.
3	During my last business trip to Canada I met my friend Ben from college.
4	As a scientist from the UK, you can be proud!
4	Four days ago, I started my blog. Stay tuned for more content.
4	2004 will be a great year for science and for my career as a biologist.
5	Did you know that Pisces is the last constellation of the zodiac.
6	Rainy weather again here in the UK. I hope you all have a good day!

Motivating Example

Structured Data

A^* | Relational Attributes $A_1, ..., A_5$

id	gender	age	topic	sign	date
1	male	[36-37]	(Education, Banking)	(Aries, Pisces)	2004-05
1	male	[36-37]	(Education, Banking)	(Aries, Pisces)	2004-05
2	male	[24-29]	(Student, indUnk)	(Leo, Pisces)	[2004-2005]
3	male	[36-37]	(Education, Banking)	(Aries, Pisces)	2004-05
4	female	[24-27]	Science	Aries	2004
4	female	[24-27]	Science	Aries	2004
4	female	[24-27]	Science	Aries	2004
5	male	[24-29]	(Student, indUnk)	(Leo, Pisces)	[2004-2005]
6	female	[24-27]	Science	Aries	2004

- Apply 2-anonymity for structured part

Unstructured Data

A^* | Textual Attribute X

id	text
1	My name is Pedro, I'm a 36 years old engineer from Mexico.
1	A quick follow up: I will post updates about my education in more detail.
2	I will start working for a big tech company as an engineer.
3	During my last business trip to Canada I met my friend Ben from college.
4	As a scientist from the UK, you can be proud!
4	Four days ago, I started my blog. Stay tuned for more content.
4	2004 will be a great year for science and for my career as a biologist.
5	Did you know that Pisces is the last constellation of the zodiac.
6	Rainy weather again here in the UK. I hope you all have a good day!

A^* direct-identifying relational attribute
A_i quasi-identifying relational attribute
X textual attribute containing sensitive terms

Motivating Example

Structured Data

A^*	Relational Attributes $A_1, ..., A_5$				
id	**gender**	**age**	**topic**	**sign**	**date**
1	male	[36-37]	(Education, Banking)	(Aries, Pisces)	2004-05
1	male	[36-37]	(Education, Banking)	(Aries, Pisces)	2004-05
2	male	[24-29]	(Student, indUnk)	(Leo, Pisces)	[2004-2005]
3	male	[36-37]	(Education, Banking)	(Aries, Pisces)	2004-05
4	female	[24-27]	Science	Aries	2004
4	female	[24-27]	Science	Aries	2004
4	female	[24-27]	Science	Aries	2004
5	male	[24-29]	(Student, indUnk)	(Leo, Pisces)	[2004-2005]
6	female	[24-27]	Science	Aries	2004

- Anonymize textual part by suppressing sensitive terms

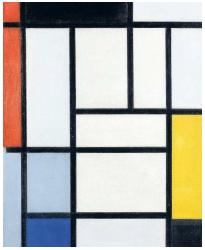
Unstructured Data

A^*	Textual Attribute X
id	**text**
1	My name is ~~Pedro~~, I'm a ~~36 years old engineer~~ from ~~Mexico~~.
1	A quick follow up: I will post updates about my education in more detail.
2	I will start working for a big tech company as an ~~engineer~~.
3	During my last business trip to ~~Canada~~ I met my friend ~~Ben~~ from college.
4	As a scientist from the ~~UK~~, you can be proud!
4	~~Four days ago~~, I started my blog. Stay tuned for more content.
4	~~2004~~ will be a great year for science and for my career as a ~~biologist~~.
5	Did you know that ~~Pisces~~ is the last constellation of the zodiac.
6	Rainy weather again here in the ~~UK~~. I hope you all have a good day!

A^* direct-identifying relational attribute
A_i quasi-identifying relational attribute
X textual attribute containing sensitive terms

Structured Data

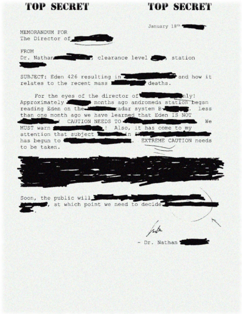
- k-anonymity as fundamental framework by Sweeney [15]
- Work on efficient implementations (e.g. Mondrian by LeFevre et. al [11])
- Work from Nergiz et. al [13] and Gong et. al [6] transferred k-anonymity to a multi-relational setting
- He and Naughton [8] adapted k-anonymity for set-valued data



Credit: https://www.tagesspiegel.de/

Textual Data

- Majority of work uses Named Entity Recognition (NER) to detect sensitive terms [3, 9, 12, 16]
- Khan et. al [10] showed that transformer-based models achieve good recall for NER tasks
- Johnson et. al [9] already used transformer models for de-identification of texts
- Liu et. al [12] emphasize a combined approach using language models and rules



TOP SECRET TOP SECRET

January 18th [redacted]

MEMORANDUM FOR
The Director of [redacted]

FROM
Dr. Nathan [redacted], clearance level [redacted], station [redacted]

SUBJECT: Eden 426 resulting in [redacted] and how it relates to the recent mass [redacted] deaths.

For the eyes of the director of [redacted]! Approximately [redacted] months ago andromeda station began reading Eden on the [redacted] system B[redacted]. Less than one month ago we have learned that Eden IS NOT [redacted] CAUTION NEEDS TO [redacted]. We MUST warn [redacted]! Also, it has come to my attention that subject [redacted] in [redacted] has begun to [redacted]. EXTREME CAUTION needs to be taken.

[redacted]

Soon, the public will [redacted], at which point we need to decide [redacted].

- Dr. Nathan [redacted]

Credit: https://media.moddb.com/

Problem Statement

Given: RX-dataset with traditional relational attributes $A_1, ..., A_n$ and textual attribute X

Ultimate Goal: Transferring sensitive terms within texts to the structured world to reduce textual information loss while making privacy guarantees

Important Definitions:

- *Redundant sensitive information*: Information with same meaning appearing in at least one relational and textual attribute of same record
- *Non-redundant sensitive information*: Poses "new" information and therefore those terms are stored in new set-valued attribute X'
- *Equivalence class*: Partition P where for any two records $r, s \in P$ $(r.A_1, ..., r.A_n) = (s.A_1, ..., s.A_n)$ and $r.X' = s.X'$
- *k-anonymity*: All equivalence classes must be at least size k [15]

Anonymization Pipeline

1. Detecting sensitive terms in textual attributes using NLP libraries
2. Linking redundant sensitive information between relational and textual attributes based on string matching
3. Building a person-centric view by aggregating data based on a direct identifier
4. Partitioning of the dataset using a pre-defined strategy
5. Recoding of relational as well as textual attributes in found partitions

Anonymization Pipeline

1. Detecting sensitive terms in textual attributes using NLP libraries
2. Linking redundant sensitive information between relational and textual attributes based on string matching
3. Building a person-centric view by aggregating data based on a direct identifier
4. Partitioning of the dataset using a pre-defined strategy
5. Recoding of relational as well as textual attributes in found partitions

Modified Mondrian Partitioning

- Recursive strategy by LeFevre et. al [11]
- Median-based splitting of partition in two sub-partitions
  - Non-numerical attributes: Sort and split by middle element
- Result is a set of partitions with size $|P| \geq k$
- **Addition:** Weight parameter λ to tune balance between relational and textual attributes

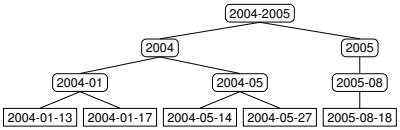
Algorithm 1: *Modified Mondrian Partitioning* - Greedy strict top-down partitioning for relational attributes adapted from [11].

Input : Partition P, relational weight λ
Output: Set of partitions with size of at least k

```
1  Function mondrian_partitioning(P, λ):
2      if |P| < 2k then
3          return P
4      end
5      else
6          A = next_attribute(λ)
7          F = frequency_set(P, A)
8          P_l = (r ∈ P | r.A < find_median(F))
9          P_r = P \ P_l
10         return mondrian_partitioning(P_l) ∪
               mondrian_partitioning(P_r)
11     end
```

Recoding - Relational Attributes

- **Recoding:** Find a single value as a replacement for multiple (probably) different values
- We use local recoding to favor utility
- For numerical and date attributes, we generate generalized values automatically
- Categorical attributes get grouped as sets



Domain Generalization Hierarchy adapted from El Emam et. al [5]

Recoding - Textual Attributes

id	gender	age	topic	sign	date	text
1	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	My name is Pedro, I'm a 36 years old engineer from Mexico.
1	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	A quick follow up: I will post updates about my education in more detail.
2	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	I will start working for a big tech company as an engineer.
3	male	[29-37]	(indUnk,Banking)	Pisces	2004-05	During my last business trip to Canada I met my friend Ben from college.
4	female	[24-27]	Science	Aries	2004	As a scientist from the UK, you can be proud!
4	female	[24-27]	Science	Aries	2004	Four days ago, I started my blog. Stay tuned for more content.
4	female	[24-27]	Science	Aries	2004	2004 will be a great year for science and for my career as a biologist.
5	male	[29-37]	(indUnk,Banking)	Pisces	2004-05	Did you know that Pisces is the last constellation of the zodiac.
6	female	[24-27]	Science	Aries	2004	Rainy weather again here in the UK. I hope you all have a good day!

1. Take recoded relational attributes as basis

Recoding - Textual Attributes

id	gender	age	topic	sign	date	text
1	male	**[24-36]**	(Student,Education)	(Leo,Aries)	[2004-2005]	My name is Pedro, I'm a **[24-36]** years old engineer from Mexico.
1	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	A quick follow up: I will post updates about my education in more detail.
2	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	I will start working for a big tech company as an engineer.
3	male	[29-37]	(indUnk,Banking)	Pisces	2004-05	During my last business trip to Canada I met my friend Ben from college.
4	female	[24-27]	Science	Aries	2004	As a scientist from the UK, you can be proud!
4	female	[24-27]	Science	Aries	2004	Four days ago, I started my blog. Stay tuned for more content.
4	female	[24-27]	**Science**	Aries	**2004**	**2004** will be a great year for **science** and for my career as a biologist.
5	male	[29-37]	(indUnk,Banking)	**Pisces**	2004-05	Did you know that **Pisces** is the last constellation of the zodiac.
6	female	[24-27]	Science	Aries	2004	Rainy weather again here in the UK. I hope you all have a good day!

1. Take recoded relational attributes as basis
2. Recode **redundant sensitive information** using replacements from linked relational attributes

Recoding - Textual Attributes

id	gender	age	topic	sign	date	text
1	male	**[24-36]**	(Student,Education)	(Leo,Aries)	[2004-2005]	My name is **person**, I'm a **[24-36]** years old **engineer** from **location**.
1	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	A quick follow up: I will post updates about my education in more detail.
2	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	I will start working for a big tech company as an **engineer**.
3	male	[29-37]	(indUnk,Banking)	Pisces	2004-05	During my last business trip to **location** I met my friend **person** from college.
4	female	[24-27]	Science	Aries	2004	As a **job** from the **UK**, you can be proud!
4	female	[24-27]	Science	Aries	2004	**Date**, I started my blog. Stay tuned for more content.
4	female	[24-27]	**Science**	Aries	**2004**	**2004** will be a great year for **science** and for my career as a **job**.
5	male	[29-37]	(indUnk,Banking)	**Pisces**	2004-05	Did you know that **Pisces** is the last constellation of the zodiac.
6	female	[24-27]	Science	Aries	2004	Rainy weather again here in the **UK**. I hope you all have a good day!

1. Take recoded relational attributes as basis
2. Recode **redundant sensitive information** using replacements from linked relational attributes
3. Recode **non-redundant sensitive information**: Sensitive terms appearing in all records of a partition can **stay**, others will be **suppressed** with their entity types

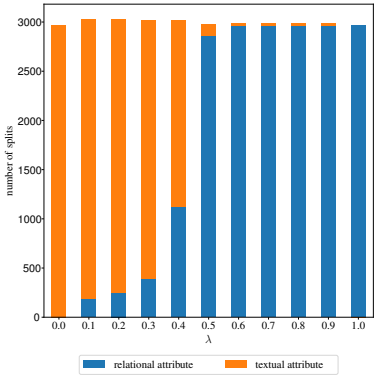
Experimental Apparatus

- Implemented anonymization pipeline in Python
- spaCy and its new transformer-based model is used to analyze textual attributes
- Run experiments on
  - *Blog Authorship Corpus*: 681,260 blog posts of 19,319 distinct bloggers [14]
  - *Hotel Reviews Dataset*: 512,126 reviews for 1,475 distinct hotels [1]
- Anonymize dataset by varying
  - k (2, 3, 4, 5, 10, 20, 50)
  - partitioning strategy and parameters (λ)
  - considered entity types (all entity types vs. only locations)
- Measure
  - statistics on split decisions and partitions
  - information loss using Normalized Certainty Penalty (NCP)

Experiment Results - Partitions

Blog Authorship Corpus, all entities

- Without modifying the partition decisions ($\lambda = 0.5$), relational attributes are favored
- λ is able to control splitting decisions for $\lambda < 0.5$
- Same partition layout for $0.6 \leq \lambda \leq 0.9$

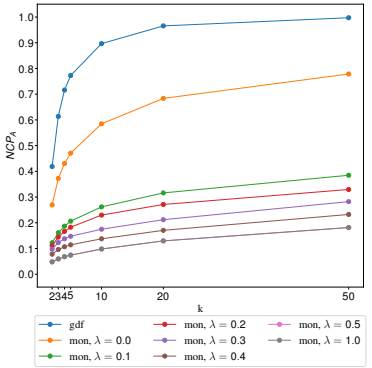


Distribution of partition splits using Mondrian for k = 5

Experiment Results - Information Loss

Blog Authorship Corpus, all entities

- Increasing k results in more information loss
- Relational information loss increases with decreasing λ

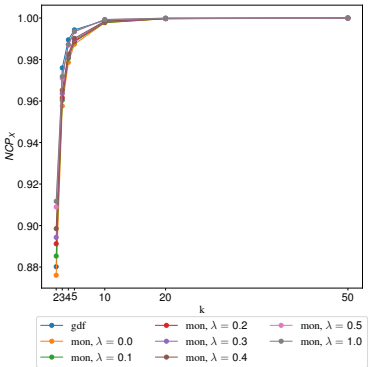


Relational Information Loss NCP_A

Experiment Results - Information Loss

Blog Authorship Corpus, all entities

- Increasing k results in more information loss
- Relational information loss increases with decreasing λ
- For small k → textual information loss less than 1
- λ can slightly control textual information loss

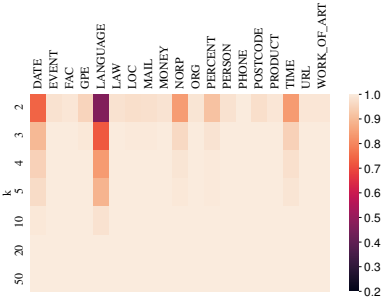


Textual Information Loss NCP_X

Experiment Results - Information Loss

Blog Authorship Corpus, all entities

- Increasing k results in more information loss
- Relational information loss increases with decreasing λ
- For small k → textual information loss less than 1
- λ can slightly control textual information loss
- LANGUAGE entities can be better preserved



Detailed textual information loss for $\lambda = 0.2$

Discussion

- k-anonymity applicable on texts by transferring the task of anonymizing sensitive terms to a structured problem
- Tuning Mondrian crucial to cope with heterogeneity of sensitive terms
- Textual information loss can be reduced under our k-anonymity model
- Over-anonymization in case of different terms with same meaning
  - "London" vs. "the capital of the UK"
- Under-anonymization if sensitive terms have different context
  - "I love London" vs. "I live in London"
- Identity disclosure might still be possible using authorship identification on texts

Conclusion and Future Work

- Combined anonymization approach achieved using k-anonymity
- Tuning and prioritization possible using λ in Mondrian partitioning
- Framework is applicable for variable datasets
- Experiments indicate that sensitive entities can be preserved

What's next?

- Abu-Khzam et. al [2]: Clustering algorithms with lower boundary on cluster size
- Hassanzadeh et. al [7]: Methods on finding non-trivial links within data
- Dwork [4]: Differential private anonymization techniques

Thank you!

References I

[1] 515k hotel reviews data in europe.
https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe.
Accessed: 2020-09-15.

[2] Abu-Khzam, F. N., Bazgan, C., Casel, K., and Fernau, H.
Clustering with Lower-Bounded Sizes: A General Graph-Theoretic Framework.
Algorithmica 80, 9 (2018), 2517–2550.

[3] Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P.
De-identification of patient notes with recurrent neural networks.
Journal of the American Medical Informatics Association 24, 3 (2017), 596–606.

[4] Dwork, C.
Differential Privacy.
In *Automata, Languages and Programming, 33rd International Colloquium* (Venice, Italy, 2006), vol. 4052, Springer, pp. 1–12.

[5] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., and Bottomley, J.
A Globally Optimal k-Anonymity Method for the De-Identification of Health Data.
Journal of the American Medical Informatics Association 16, 5 (2009), 670–682.

[6] Gong, Q., Luo, J., Yang, M., Ni, W., and Li, X. B.
Anonymizing 1:M microdata with high utility.
Knowledge-Based Systems 115 (2017), 15–26.

References II

[7] Hassanzadeh, O., Lim, L., Kementsietsidis, A., and Wang, M.
A declarative framework for semantic link discovery over relational data.
In *WWW'09 - Proceedings of the 18th International World Wide Web Conference* (Madrid, Spain, 2009), ACM, pp. 1101–1102.

[8] He, Y., and Naughton, J. F.
Anonymization of Set-Valued Data via Top-Down, Local Generalization.
Proceedings of the VLDB Endowment 2, 1 (2009), 934–945.

[9] Johnson, A. E. W., Bulgarelli, L., and Pollard, T. J.
Deidentification of free-text medical records using pre-trained bidirectional transformers.
In *Proceedings of the ACM Conference on Health, Inference, and Learning* (Toronto, Ontario, Canada, 2020), ACM, pp. 214–221.

[10] Khan, M. R., Ziyadi, M., and AbdelHady, M.
MT-BioNER: Multi-task Learning for Biomedical Named Entity Recognition using Deep Bidirectional Transformers.
2020.

[11] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R.
Mondrian Multidimensional K-Anonymity.
In *Proceedings of the 22nd International Conference on Data Engineering* (Atlanta, GA, USA, 2006), IEEE, pp. 25–25.

[12] Liu, Z., Tang, B., Wang, X., and Chen, Q.
De-identification of clinical notes via recurrent neural network and conditional random field.
Journal of Biomedical Informatics 75 (2017), 34–42.

References III

[13] Nergiz, M. E., Clifton, C., and Nergiz, A. E.
MultiRelational k-Anonymity.
In *Proceedings of the 23rd International Conference on Data Engineering* (Istanbul, Turkey, 2007), vol. 21, IEEE, pp. 1417–1421.

[14] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J.
Effects of age and gender on blogging.
In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium* (Stanford, California, USA, 2006), AAAI, pp. 199–205.

[15] Sweeney, L.
k-Anonymity: A Model for Protecting Privacy.
International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 5 (2002), 557–570.

[16] Trienes, J., Trieschnigg, D., Seifert, C., and Hiemstra, D.
Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records.
In *Proceedings of the ACM WSDM 2020 Health Search and Data Mining Workshop* (Houston, TX, USA, 2020), vol. 2551, CEUR-WS.org, pp. 3–11.

Anonymization Framework