Data Anonymizer
Hack the Threat 2023

Team and high-level overview

Team Members

- Michelle
- Ziggy
- Francesco
- Kai
- Max
- Seanyoon

Task

Create a automated process that can take in sensitive data and make it less sensitive. Ideally a user can provide a file, have anonymization done automatically, and then their file can be returned in the same format and shape.

Link to project (see readme for web app link): https://github.com/kmaurinjones/2hack2furious

WHAT IS "ANONYMOUS DATA"?

Goal and Workflow

Why "2"-Anonymous?

Ease of use & accessibility

The original paper:
"Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"

Substitute or Redact?

Dealing with different data types
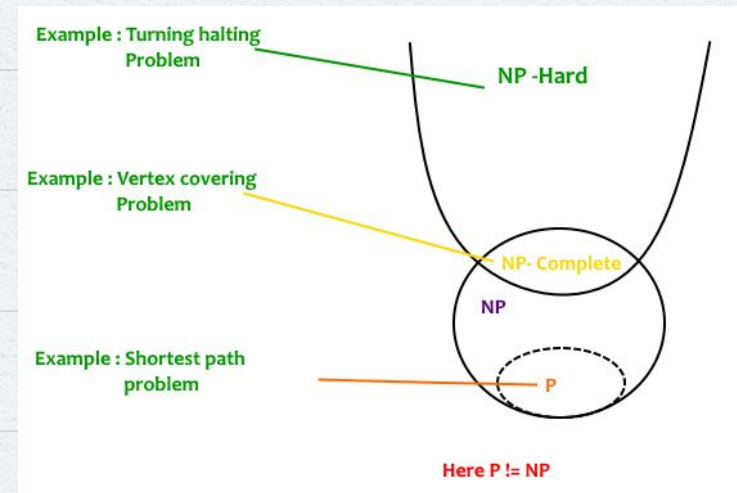
4

K-anonymity is not foolproof!

A determined person may still be able to identify individuals if they have more than k pieces of information.
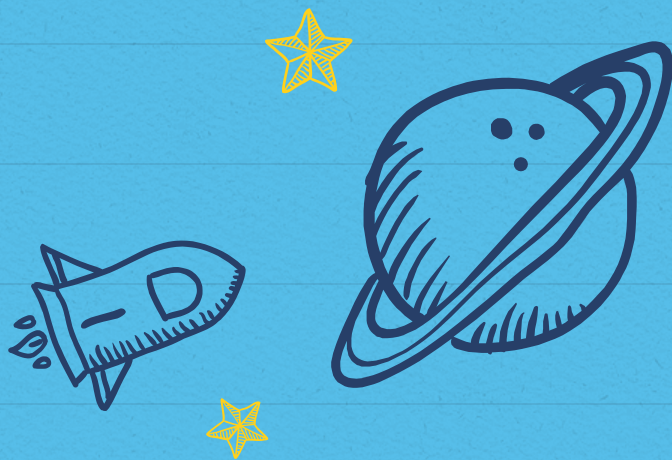
An NP-hard Problem
On the Complexity of Optimal K-Anonymity

- To achieve k-anonymity, we need to group individuals in a way that minimizes the amount of information that could potentially identify them, while still satisfying the requirement that each group has at least k members.

- This requires exploring all possible combinations of individuals and grouping them in different ways to see which groupings meet the k-anonymity criteria.

- This problem is "NP-Hard" since it can't be solved optimally and a solution cannot be verified in polynomial time with current algorithms.

Example : Turning halting Problem — NP -Hard

Example : Vertex covering Problem — NP- Complete

NP

Example : Shortest path problem — P

Here P != NP

Our Web app

Hosted on HuggingFace

Benefits: Interactivity, functionality... and it's free!

The Workflow

Upload CSV

Do automatic substitutions and redactions

Show data preview to user

1

3

5

2

4

6

Pre-process Data

Run k-anonymity analysis

Return data to user

Our Vision

2data2anonymous

Upload Data

Auto clean:
✔ **Remove duplicates**
✔ **Remove missing**
✔ **Clean formatting**

Auto anonymize:
✔ **Delete columns**
✔ **Substitute columns**

Fine tuning:
· **Bin size**
· **Normalisation**

Disclaimer:
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ut nunc augue. Morbi ut neque augue. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia curae; Ut sed orci vulputate, suscipit ligula sit amet, convallis enim. Donec metus lectus, posuere ut elit ac, pulvinar mollis orci. Donec viverra pulvinar diam, a porttitor felis pharetra a.

Franceso	18	UBC	0.72	118 Green Dr, ...
Sean	27	SFU	1.45	243 Blue Rd, ...
Kai	64	UBC	2.32	19 Blue Rd, ...
Max	53	UTexas	0.18	8 Orange Ln, ...
Michelle	25	UVic	1.74	117 Red Rd, ...
Je'aime	19	SFU	4.39	97 Blue Rd, ...
Ricardo	24	UVic	5.25	18 Orange Ln, ...
Sasmita	72	UBC	1.34	13 Green Dr, ...
Ziggy	84	UTexas	1.27	4 Red Rd, ...
Charlie	39	UTexas	0.29	6 Red Rd, ...

★	<21	UBC	0.72	Green Dr, ...
★	25-30	SFU	1.45	Blue Rd, ...
★	60-65	UBC	2.32	Blue Rd, ...
★	50-55	UTexas	0.18	Orange Ln, ...
★	25-30	UVic	1.74	Red Rd, ...
★	<21	SFU	4.39	Blue Rd, ...
★	20-25	UVic	5.25	Orange Ln, ...
★	70-75	UBC	1.34	Green Dr, ...
★	80-85	UTexas	1.27	Red Rd, ...
★	35-40	UTexas	0.29	Red Rd, ...

7 score

Download

miro

Our Result

2anonymity

Clean and anonymize data

Drag and drop file here
Limit 5MB per file • CSV, JSON, XLSX

[Browse files]

📄 HRDataset_v14.csv   ✕
75.1KB

Data loading options:
☑ Remove duplicate rows
☐ Remove rows with missing values

Anonymizing options:

Categorical Variable Threshold
50
2   200

Bin Size
20
2   200

Redaction strength
Low ▼

Download options:

File type
.csv ▼

[Download]

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	Salary
0	Adinolfi, Wilson K	10026	0	0	1	1	5	4	0	62506
1	Ait Sidi, Karthikeyan	10084	1	1	1	5	3	3	0	104437
2	Akinkuolie, Sarah	10196	1	1	0	5	5	3	0	64955
3	Alagbe,Trina	10088	1	1	0	1	5	3	0	64991
4	Anderson, Carol	10069	0	2	0	5	5	3	0	50825
5	Anderson, Linda	10002	0	0	0	1	5	4	0	57568
6	Andreola, Colby	10194	0	0	0	1	4	3	0	95660
7	Athwal, Sam	10062	0	4	1	1	5	3	0	59365
8	Bachiochi, Linda	10114	0	0	0	3	5	3	1	47837
9	Bacong, Alejandro	10250	0	2	1	1	3	3	0	50178

	Employee_Name	EmpID		MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	S
0	None	10022	10042	0	0	1	1	5	4	0	
1	None	10064	10084	1	1	1	5	3	3	0	
2	None	10190	10210	1	1	0	5	5	3	0	
3	None	10085	10105	1	1	1	1	5	3	0	
4	None	10064	10084	0	2	0	5	5	3	0	
5	None	10001	10021	0	0	0	1	5	4	0	
6	None	10190	10210	0	0	0	1	4	3	0	
7	None	10043	10063	0	4	1	1	5	3	0	
8	None	10106	10126	0	0	0	3	5	3	1	
9	None	10232	10251	0	2	1	1	3	3	0	

Thank You!

Special thanks to Je'aime & Charlie!