Amnesia

Data anonymization made easy

https://amnesia.openaire.eu

Manolis Terrovitis
mter@imis.athena-innovation.gr
http://web.imsi.athenarc.gr/~mter/
Research Center Athena, IMSI

Data anonymization?

- Data anonymization facilitates the publication of micro data(vs. aggregated macrodata) , e.g., data used in scientific research

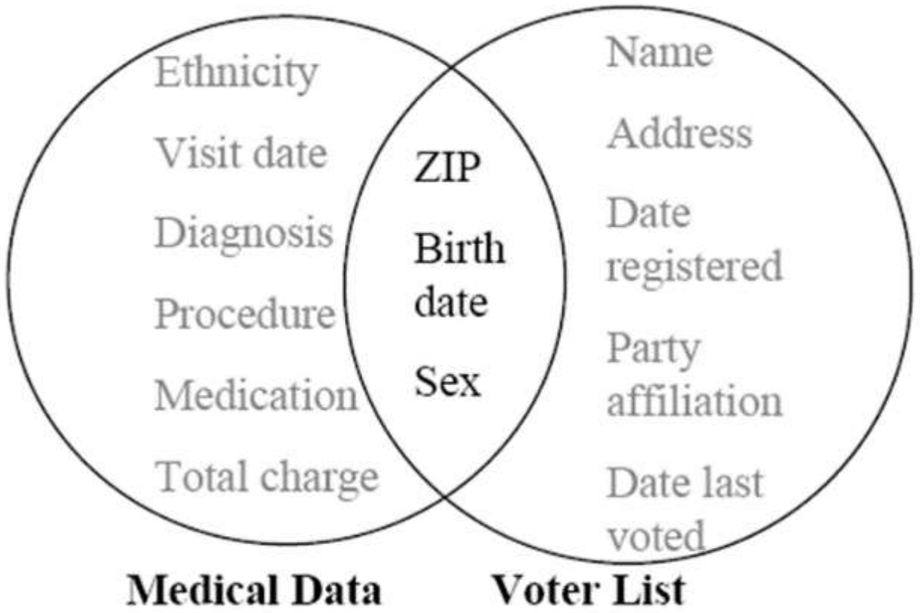
- Micro data often reveal important private information, e.g., the medical condition of a person
  - Individuals are afraid to provide their data
  - Companies are afraid to share data with experts
  - GDPR makes a strict protection scheme obligatory

- The aim of anonymization methods is to allow sharing such data, without compromising the privacy of the users.

Data anonymization and Amnesia

- Data anonymization
  - Removal of direct identifiers, e.g., Names, SSN etc
  - Removal of infrequent combinations of quasi-identifiers, e.g., unique combinations of birth dates and zipcodes
  - Infrequent combinations are removed through generalization, e.g., birth date 14/01/1977 becomes **/**/1977
- Amnesia is a scalable anonymization tool
  - It offers several versions of k-anonymity
  - It allows the user to select and customize possible solutions
  - It offers graphical tools that allow the user to analyze the anonymized dataset
  - It is scalable and uses all available CPU cores in the anonymization process

Link attacks



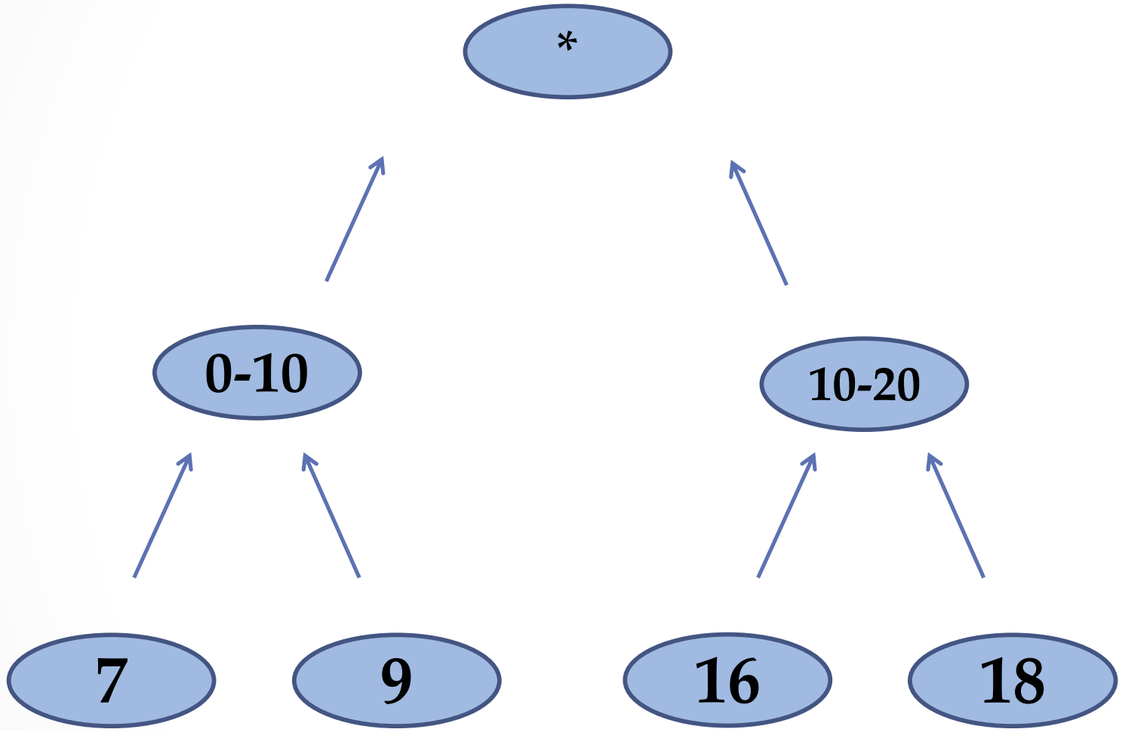
k-anonymity

- Each entry becomes indistinguishable from other *k*-1 entries
  - *k*-anonymity is achieved through suppression and generalization

id	Zipcode	Age	National.	Disease
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

~~id~~	Zipcode	Age	National.	Disease
~~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

Generalization Hierarchy



Structural information

- We need to anonymize all relevant information about a person, not just a tuple
- Information tends to gather over time
- Information is linked through semantic properties, it's schema is irrelevant
- Personal data tend to accumulate over time
- Research focuses on simple data and complicated guaranties but real world has complex data and requires simple guaranties

Limits of *k*-anonymity

	Fruits	Meat	Vegetables	Fish
Vassilis	X	X		
Manolis	X	X	X	
Eleni			X	
Maria		X	X	
Kostas	X			X

	Food
Vassilis	X
Manolis	X
Eleni	X
Maria	X
Kostas	X

- 2-anonymous

k^m-anonymity

	Fruits	Meat	Vegetables	Fish
Vassilis	X	X		
Manolis	X	X	X	
Eleni			X	
Maria		X	X	
Kostas	X			X

	Fruits	Meat	Other food
Vassilis	X	X	
Manolis	X	X	X
Eleni			X
Maria		X	X
Kostas	X		X

- 2^2-anonymous
- Any combination of *m* items will not appear less than *k* times

Strengths and Weaknesses

- Strengths
  - Simple to understand
    - Can be the basis for consent
  - Close to previous and existing legal definitions
  - Low information loss
  - Customizable by non-experts
- Weaknesses
  - Not very strict
  - Does not take into account sensitive values

Anonymization challenges

- Anonymization techniques have not been tested in practice extensively
  - Mapping the social notion of privacy to technical notions is not easy
- Data utility has not been studied extensively in research
  - Few artificial information loss measures
- Data utility is difficult to estimate in practice
  - Different applications have different needs
  - No easy to quantify the loss of information

Amensia

- Amnesia is a data anonymization tool developed by Research Center Athena
- Amnesia is build with Java and Javascript
- *k*-anonymity and k^m-anonymity
- Tuples and set-values
- Visual tools
  - Estimating data utility
  - Building hierarchies
  - Customizing anonymization solutions

Amnesia status

- Amnesia is available as a public beta version at
  - https://amnesia.openaire.eu
- On-line version is for demonstration and testing purposes mostly
- Sensitive data can be anonymized locally by downloading the application
  - Security
  - Scalability
- We are in process of adjusting it to health data

Amensia Challenges

Is it easy to use by data owners?

- Give us feedback!!
  - amnesia-helpdesk@imis.athena-innovation.gr
- Can it anonymize your data?
  - Let us know about your use case
  - Ask us for help

Are anoymized data useful?

- We need feedback for data analysis
  - Let us know if you have shared anonymized results
- Please contact us with your needs

Next steps

Work on the feedback

- Improve user experience
- Add support for specific domain data
- Fix bugs!

More features

- New algorithms
  - Additional privacy guaranties
  - More data types
- Better scaling capabilities
  - Disk based solutions
  - More efficient memory usage

Thank you!

HTTPS://AMNESIA.OPENAIRE.EU/