Προστασία της ιδιωτικότητας στην δημοσίευση δεδομένων

ΜΑΝΩΛΗΣ ΤΕΡΡΟΒΙΤΗΣ

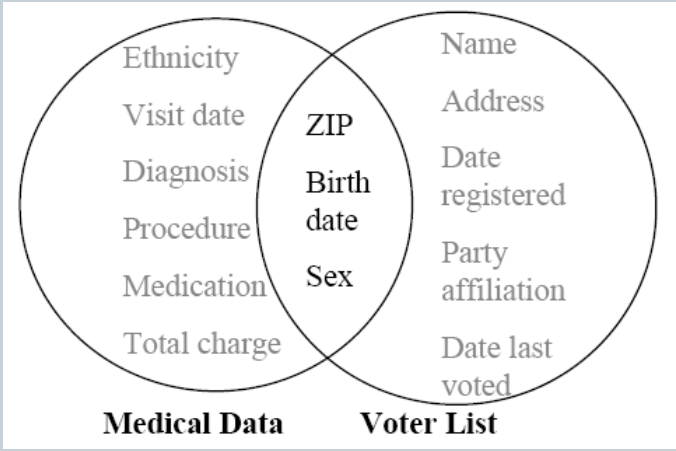
ΙΝΣΤΙΤΟΥΤΟ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΕΡΕΥΝΗΤΙΚΟ ΚΕΝΤΡΟ ΑΘΗΝΑ
MTER@IMIS.ATHENA-INNOVATION.GR
HTTP://WEB.IMIS.ATHENA-INNOVATION.GR/~MTER/

Το πρόβλημα

- Πολλοί οργανισμοί και εταιρείες δημοσιεύουν πρωτογενή δεδομένα (microdata)
- Αφαιρούν τα άμεσα αναγνωριστικά (Όνομα, ΑΦΜ κτλ)
- Παραμένει ο κίνδυνος των επιθέσεων συσχέτισης!
  - Άλλες, γνωστές πηγές πληροφορίας (π.χ. τηλεφωνικοί κατάλογοι) μπορούν να οδηγήσουν στην πραγματική ταυτότητα των χρηστών

Ένα παράδειγμα

- Ερευνητές έδειξαν ότι ήταν δυνατόν να ανακαλυφθεί το ιατρικό ιστορικό των κυβερνήτη της Μασαχουσέτης από πηγές που δεν τον κατονόμαζαν
  - Δεδομένα νοσηλείας χωρίς ονόματα
  - Εκλογικοί κατάλογοι
  - Συνδυάζοντας τα παραπάνω:
    - 6 άνθρωποι γεννημένοι την ίδια μέρα
    - 3 άντρες
    - 1 με τον ίδιο ΤΚ



- 87% του πληθυσμού τον ΗΠΑ ήταν μοναδικοί με βάση τον ΤΚ, το γένος και την ημερομηνία γέννησης

Και ένα ακόμη

- Η AOL δημοσιεύει τον Αύγουστο ανωνυμοποιημένα δεδομένα για 21Μ χρήστες
- Τα δεδομένα θα επέτρεπαν σε δικτυακές εταιρείες να έχουν μία καλή εικόνα των χρηστών
- Ο χρήστης 4417749 είχε πολλές ερωτήσεις για συγκεκριμένη τοποθεσία και θέμα
- Η NY Times έψαξαν τα δεδομένα…
- Και ανακάλυψαν ότι ο χρήστης 4417749 είναι η κα Arnold, 62, που έψαχνε για φάρμακα, σκύλους και μέλη της οικογένειας της…

Βασική Ιδέα

Μετασχηματισμός των αρχικών δεδομένων σε μία μορφή που αντιμετωπίζει τους κινδύνους (ανωνυμοποίηση)

- Ποιους κινδύνους θα αντιμετωπίσουμε;
- Ποιους μετασχηματισμούς θα εφαρμόσουμε;
- Πόση πληροφορία θα χάσουμε;

Η αποτελεσματική αντιμετώπιση των κινδύνων και ο περιορισμός της απώλειας πληροφορίας είναι ανταγωνιστικοί παράγοντες

Οι κίνδυνοι

- Αποκάλυψη συμμετοχής
  - Όταν και μόνο η συμμετοχή στα δημοσιευμένα δεδομένα είναι ευαίσθητη πληροφορία
- Αναγνώριση
  - Όταν εντοπίζουμε μία εγγραφή που αναφέρεται σε ένα πρόσωπο
- Συσχέτιση
  - Όταν συσχετίζουμε μία ευαίσθητη τιμή με ένα πρόσωπο

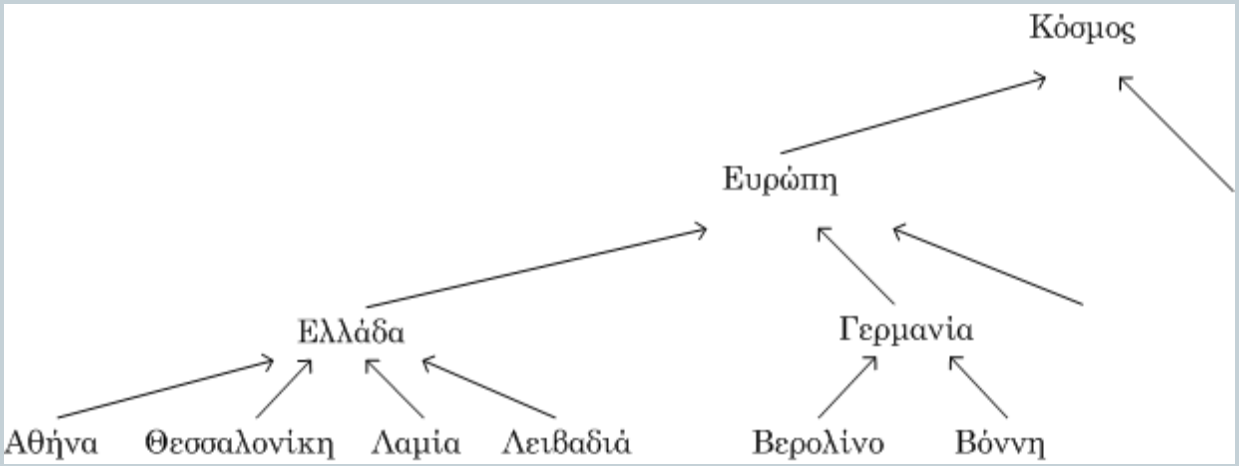
Μετασχηματισμοί

- Απόκρυψη
- Εναλλαγή Τιμών
- Αποσυσχέτιση
- Γενίκευση

Ηλικία	Πόλη
28	Αθήνα
33	Θεσσαλονίκη
21	Λαμία
24	Λειβαδιά
37	Βερολίνο
36	Βόννη
35	Βερολινο

Ηλικία	Χώρα
[25-35]	Ελλάδα
[25-35]	Ελλάδα
[20-25]	Ελλάδα
[20-25]	Ελλάδα
[35-40]	Γερμανία
[35-40]	Γερμανία
[35-40]	Γερμανία

Μετασχηματισμοί - συνέχεια



k-Ανωνυμία

- Προτάθηκε το 2001 από την Sweeney
- Διαισθητικά κρύβει κάθε πρόσωπο μέσα σε άλλες **εγγραφές**

k-Ανωνυμία - Ένα Παράδειγμα

Τα εργαλεία

- Γενίκευση
  - Δημοσίευση γενικότερων τιμών
- απόκρυψη
  - Πλειάδων, με πολύ μοναδικές τιμές (outliers)
  - Συχνά υπάρχει κάποιο όριο

Πρωτογενή δεδομένα

Ημ. Γεν.	Γένος	ΤΚ
21/1/79	Άντρας	53715
10/1/79	Γυναίκα	55410
1/10/44	Γυναίκα	90210
21/2/83	Άντρας	02274
19/4/82	Άντρας	02237

2-ανώνυμα δεδομένα

	Ημ. Γεν.	Γένος	ΤΚ
Ομάδα 1	*/1/79	πρόσωπο	5****
	*/1/79	πρόσωπο	5****
απόκρυψη	1/10/44	Γυναίκα	90210
Ομάδα 2	*/*/8*	Άντρας	022**
	//8*	Άντρας	022**

k^m-ανωνυμία

	Ελεύθερο επάγγελμα	Εισοδήματα μισθωτών	Ενοίκια	Μερίσματα
Βασίλης	X	X		
Μανώλης	X	X	X	
Ελένη			X	
Μαρία		X	X	
Κώστας	X			X

	Ελεύθερο επάγγελμα	Εισοδήματα μισθωτών	Άλλα εισοδήματα
Βασίλης	X	X	
Μανώλης	X	X	X
Ελένη			X
Μαρία		X	X
Κώστας	X		X

- 2^2-ανώνυμα δεδομένα

Ανακεφαλαίωση

- Ή προστασία της ιδιωτικότητας είναι πολύ σημαντικό πρόβλημα όταν θέλουμε να δημοσιεύσουμε δεδομένα χωρίς να αποκαλύψουμε ευαίσθητη πληροφορία για τους χρήστες
- Στην διαδικασία επίλυσης του προβλήματος πρέπει να αντιμετωπίσουμε τα εξής θέματα:
  - Ποιοι είναι οι κίνδυνοι για την ιδιωτικότητα στην πληροφορία που δημοσιεύουμε;
  - Ποιοι είναι οι ενδεδειγμένοι μετασχηματισμοί;
  - Πως θα μετρήσουμε την απώλεια της πληροφορίας
- Ποιότητα πληροφορίας vs επίπεδο προστασίας