Προστασία της ιδιωτικότητας - Ανωνυμοποίηση προσωπικών δεδομένων

Info-Sec-Lab

Η εποχή των «μεγάλων δεδομένων» (Big Data)

- **V**olume
  - Ο όγκος των ψηφιακών δεδομένων στο Διαδίκτυο αγγίζει τα 9.5 δισεκατομμύρια petabytes (9.5 x 10^{24} bytes) για το 2015, επαυξημένος κατά 3 δισεκατομμύρια petabytes σε σχέση με το 2014
  - (Πηγή: M. Meeker, 2016 Internet Trends, http://www.kpcb.com/blog/2016-internet-trends-report)
- **V**elocity
  - Όχι πλέον μόνο στατικά αλλά και δυναμικά δεδομένα (εκατομμύρια «κλικ» των χρηστών ανά δευτερόλεπτο, τα οποία «αντανακλούν» π.χ., τις συνήθειές τους)
- **V**ariety
  - Αριθμοί, εικόνες, ήχος, βίντεο, 3D δεδομένα, γεωχωρικά δεδομένα, δεδομένα από κοινωνικά δίκτυα,…

Πηγή: J. Domingo-Ferer, "Directions in Big Data Anonymization", 2016

Απειλές για την ιδιωτικότητα

- Ευχερής δυνατότητα συγκέντρωσης και αξιοποίησης πολλών πληροφοριών από πολλές διαφορετικές πηγές μπορεί να επιφέρει ταυτοποίηση/αναγνώριση προσώπου από φαινομενικά «ανώνυμα» δεδομένα
- Αυτό με τη σειρά του μπορεί να οδηγήσει σε:
  - Αποκάλυψη ευαίσθητων πληροφοριών
  - Χρήση των προσωπικών του δεδομένων για άλλους σκοπούς, ενδεχομένως μη συμβατούς με τους αρχικούς
  - Όχι διαφάνεια στην επεξεργασία των προσωπικών δεδομένων
- Παράδειγμα: Εταιρεία ανέπτυξε μοντέλο για πρόγνωση εγκυμοσύνης - παρατηρώντας, π.χ., τις διαδικτυακές παραγγελίες προϊόντων – και «μάντεψε» εγκυμοσύνη έφηβης, καθώς και το μήνα εγκυμοσύνης, πριν το μάθουν οι γονείς της
  - Πηγή: C. Duhigg (2012) How companies learn your secrets, New York Times Magazine
  - http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
- Τελικά, τι είναι ιδιωτικότητα;
  - «Ευχερέστερο να την προστατέψει κανείς, παρά να την περιγράψει»

Εισαγωγικές σκέψεις

- Η ιδιωτικότητα είναι στενά συνυφασμένη – αν και δεν ταυτίζεται – με την προστασία προσωπικών δεδομένων
- Στην εποχή των μεγάλων δεδομένων, γίνεται συχνά επεξεργασία προσωπικών δεδομένων
  - Ποια δεδομένα είναι προσωπικά;
  - Πότε είναι νόμιμη η επεξεργασία προσωπικών δεδομένων;
  - Τι τεχνικές πρέπει να εφαρμόζονται για να είναι νόμιμη η επεξεργασία;

Η έννοια της ιδιωτικότητας

- Ιδιωτικότητα (Privacy):
  - The ability of the individual to control the terms under which personal information is acquired and used. (Westin A..F., 1967)
  - A state or condition of limited access to a person, information about him, intimacies of personal identity (F. Schoeman, 1984)
  - The right to privacy is the right to be left alone (Brandeis, 1928)

- Με το θεσμικό πλαίσιο της προστασίας προσωπικών δεδομένων, τίθενται προϋποθέσεις νομιμότητας της επεξεργασίας προσωπικών δεδομένων, καθώς επίσης αναγνωρίζονται συναφή δικαιώματα και υποχρεώσεις, στο πλαίσιο προστασίας του θεμελιώδους αγαθού της ιδιωτικότητας.

Ιστορικά στοιχεία – Εισαγωγή

- 1950: Η προστασία της ιδιωτικής ζωής αναγνωρίζεται ως δικαίωμα στην ΕΣΔΑ (άρθρο 8)
  - Παν πρόσωπον δικαιούται εις τον σεβασμόν της ιδιωτικής και οικογενειακής ζωής του, της κατοικίας του και της αλληλογραφίας του (…).

- 1981: Συμβούλιο της Ευρώπης, Σύμβαση 108 για την Προστασία του Ατόμου από την Αυτοποιημένη Επεξεργασία Προσωπικών Δεδομένων
  - (…) Οι πληροφορίες προσωπικού χαρακτήρα που αποκαλύπτουν τη φυλετική προέλευση, τα πολιτικά φρονήματα, τις θρησκευτικές ή άλλες πεποιθήσεις, όπως και οι πληροφορίες προσωπικού χαρακτήρα που σχετίζονται με την υγεία ή την σεξουαλική ζωή, δεν δύνανται να αποτελέσουν αντικείμενο αυτοματοποιημένης επεξεργασίας, εάν το εσωτερικό δίκαιο δεν προβλέπει κατάλληλες εγγυήσεις. Το αυτό ισχύει για τις πληροφορίες προσωπικού χαρακτήρα που αφορούν ποινικές καταδίκες (…)

Προστασία προσωπικών δεδομένων

- 1995: Οδηγία 95/46/ΕΚ για την προστασία των φυσικών προσώπων από την επεξεργασία των προσωπικών δεδομένων τους και την ελεύθερη διακίνηση των δεδομένων (Ευρωπαϊκό Συμβούλιο και Κοινοβούλιο).
- Μία Ευρωπαϊκή Οδηγία – όπως η παραπάνω - πρέπει να ενσωματωθεί στην εθνική νομοθεσία κάθε Κράτους – Μέλους
  - Άρα, όλα τα Κράτη – Μέλη έχουν σχεδόν το ίδιο νομικό πλαίσιο, αφού το κάθε ένα έχει στην έννομη τάξη του ένα νόμο που βασίζεται στην ίδια Οδηγία
  - Υπάρχουν όμως πάντα μικρές διαφοροποιήσεις (η ίδια η Οδηγία αφήνει διάφορα θέματα ανοικτά στον εκάστοτε εθνικό νομοθέτη).
  - Στην Ελλάδα: Νόμος 2472/1997
    - Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
- 2001: Συνταγματική αναθεώρηση.
  - Άρθρο 9Α: Καθένας έχει δικαίωμα προστασίας από τη συλλογή, επεξεργασία και χρήση, ιδίως με ηλεκτρονικά μέσα, των προσωπικών του δεδομένων, όπως νόμος ορίζει. Η προστασία των προσωπικών δεδομένων διασφαλίζεται από ανεξάρτητη αρχή, που συγκροτείται και λειτουργεί, όπως νόμος ορίζει.

Τι άλλαξε στις 25 Μαΐου 2018

- Νέος Κανονισμός 2016/679 του Ευρωπαϊκού Κοινοβουλίου σε θέματα προστασίας των ατόμων για την επεξεργασία προσωπικών δεδομένων - General Data Protection Regulation (GDPR) (αντικατάσταση της Οδηγίας 95/46/ΕΚ)
  - Ενίσχυση της προστασίας προσωπικών δεδομένων
  - Εναρμόνιση βασικών κανόνων
  - Άμεση εφαρμογή σε όλα τα Κράτη Μέλη το Μάιο του 2018
  - Νέες υποχρεώσεις για τους υπευθύνους επεξεργασίας
  - Ενίσχυση των δικαιωμάτων των πολιτών
- Δεν χρειάζεται εθνική νομοθεσία – έχει άμεση εφαρμογή
  - Για λίγα επιμέρους θέματα αφήνει περιθώρια ο GDPR στον εθνικό νομοθέτη να εξειδικεύσει κάποια ζητήματα
- Άρα: μεγαλύτερος βαθμός εναρμόνισης μεταξύ των Κρατών Μελών (αφού με την Οδηγία 95/46/ΕΚ υπήρχε δυνατότητα διαφοροποίησης σε διάφορα σημεία της νομοθεσίας για το Κράτος Μέλος)

- Θα μελετηθεί σε επόμενα μαθήματα, σε σχέση με τις αλλαγές που επιφέρει στο νυν νομικό πλαίσιο
  - Στο μάθημα αυτό, θα σταθούμε στους ορισμούς των προσωπικών και των ανώνυμων δεδομένων

Προσωπικά Δεδομένα

- Προσωπικά δεδομένα (ή Δεδομένα προσωπικού χαρακτήρα):
  - κάθε πληροφορία (άμεση ή έμμεση) που αναφέρεται σε φυσικό πρόσωπο και χαρακτηρίζει το υποκείμενο από φυσική, βιολογική, ψυχολογική, οικονομική, πολιτιστική ή κοινωνική άποψη
  - Δεν λογίζονται ως προσωπικά δεδομένα τα στατιστικής φύσεως συγκεντρωτικά στοιχεία
- Στην πράξη, προσωπικά δεδομένα είναι κάθε πληροφορία που μας χαρακτηρίζει, όπως για παράδειγμα:
  - το όνομά μας
  - η διεύθυνσή μας (ταχυδρομική αλλά και ηλεκτρονική – email),
  - το τηλέφωνό μας,
  - τα ενδιαφέροντά μας,
  - οι απόψεις μας,
  - η εικόνα μας (φωτογραφία/video)
  - …..
- Είναι όμως και πολλά περισσότερα, που ίσως δεν φανταζόμαστε
  - Το ψευδώνυμό μας (nickname) σε μία διαδικτυακή υπηρεσία, ακόμα και αν δεν παραπέμπει στο πραγματικό μας ονοματεπώνυμο
  - Η IP διεύθυνση του υπολογιστή μας από τον οποίο «σερφάρουμε»
  - ….

Ευαίσθητα προσωπικά δεδομένα

- Κάποια προσωπικά δεδομένα χρήζουν ακόμα μεγαλύτερης προστασίας, γιατί εμπίπτουν στο σκληρό πυρήνα της ιδιωτικότητας

- Ευαίσθητα δεδομένα: τα δεδομένα που αφορούν σε
  - φυλετική ή εθνική προέλευση,
  - πολιτικά φρονήματα,
  - θρησκευτικές ή φιλοσοφικές πεποιθήσεις,
  - συμμετοχή σε συνδικαλιστική οργάνωση,
  - υγεία,
  - κοινωνική πρόνοια,
  - ερωτική ζωή,
  - ποινικές διώξεις ή καταδίκες,
  - στη συμμετοχή σε συναφείς με τα παραπάνω ενώσεις
- Οι ορισμοί αυτοί υπάρχουν στην Οδηγία 95/46/ΕΚ και, άρα, και στο ν. 2472/997

Ευαίσθητα προσωπικά δεδομένα (συνέχεια)

- Με τον GDPR, έχουν προστεθεί άλλες δύο κατηγορίες ευαίσθητων προσωπικών δεδομένων:
- **Γενετικά δεδομένα**
  - τα δεδομένα προσωπικού χαρακτήρα που αφορούν τα γενετικά χαρακτηριστικά φυσικού προσώπου που κληρονομήθηκαν ή αποκτήθηκαν, όπως προκύπτουν, ιδίως, από ανάλυση βιολογικού δείγματος του εν λόγω φυσικού προσώπου και τα οποία παρέχουν μοναδικές πληροφορίες σχετικά με την φυσιολογία ή την υγεία του εν λόγω φυσικού προσώπου (άρθρο 4)
    - Π.χ. δεδομένα που προκύπτουν από ανάλυση DNA, RNA κτλ.
- **Βιομετρικά δεδομένα**, εφόσον χρησιμοποιούνται για την ταυτοποίηση ατόμου, θεωρούνται πλέον ως τέτοιας ιδιαίτερης κατηγορίας δεδομένα
  - δεδομένα προσωπικού χαρακτήρα τα οποία προκύπτουν από ειδική τεχνική επεξεργασία συνδεόμενη με φυσικά, βιολογικά ή συμπεριφορικά χαρακτηριστικά φυσικού προσώπου και τα οποία επιτρέπουν ή επιβεβαιώνουν την αδιαμφισβήτητη ταυτοποίηση του εν λόγω φυσικού προσώπου, όπως εικόνες προσώπου ή δακτυλοσκοπικά δεδομένα (άρθρο 4)
    - Π.χ. Εφαρμογές facial recognition

Ανώνυμα δεδομένα

Οδηγία 95/46/ΕΚ:

Οι αρχές της προστασίας δεν εφαρμόζονται σε δεδομένα που έχουν καταστεί ανώνυμα κατά τρόπο ώστε να μην μπορεί να εξακριβωθεί πλέον η ταυτότητα του προσώπου στο οποίο αναφέρονται

- Για να διαπιστωθεί αν η ταυτότητα ενός προσώπου μπορεί να εξακριβωθεί, πρέπει να λαμβάνεται υπόψη το σύνολο των μέσων που μπορούν ευλόγως να χρησιμοποιηθούν, είτε από τον υπεύθυνο της επεξεργασίας, είτε από τρίτο, για να εξακριβωθεί η ταυτότητα του εν λόγω προσώπου

Κανονισμός (ΕΕ) 2016/679 (General Data Protection Regulation - GDPR)

- Οι αρχές της προστασίας δεν θα πρέπει να εφαρμόζονται σε ανώνυμες πληροφορίες, δηλ. πληροφορίες που δεν σχετίζονται προς ταυτοποιημένο ή ταυτοποιήσιμο πρόσωπο ή σε δεδομένα που έχουν καταστεί ανώνυμα κατά τρόπο ώστε η ταυτότητα του υποκειμένου να μην μπορεί να εξακριβωθεί

- Για να κριθεί κατά πόσον ένα φυσικό πρόσωπο είναι ταυτοποιήσιμο, θα πρέπει να λαμβάνονται υπόψη όλα τα μέσα τα οποία είναι ευλόγως πιθανό ότι θα χρησιμοποιηθούν, όπως για παράδειγμα ο διαχωρισμός του (singling out), είτε από τον υπεύθυνο επεξεργασίας είτε από τρίτο για την άμεση ή έμμεση εξακρίβωση της ταυτότητας του φυσικού προσώπου.

- Για να διαπιστωθεί κατά πόσον κάποια μέσα είναι ευλόγως πιθανό ότι θα χρησιμοποιηθούν για την εξακρίβωση της ταυτότητας του φυσικού προσώπου, θα πρέπει να λαμβάνονται υπόψη όλοι οι αντικειμενικοί παράγοντες, όπως τα έξοδα και ο χρόνος που απαιτούνται για την ταυτοποίηση, λαμβανομένων υπόψη της τεχνολογίας που είναι διαθέσιμη κατά τον χρόνο της επεξεργασίας και των εξελίξεων της τεχνολογίας.

Εφαρμόζεται η νομοθεσία στα ανώνυμα δεδομένα;

- Τόσο με την Οδηγία 95/46/ΕΚ (ν. 2472/1997), όσο και με τον GDPR, η απάντηση είναι ΟΧΙ:
  - Οι αρχές της προστασίας δεν εφαρμόζονται σε δεδομένα που έχουν καταστεί ανώνυμα κατά τρόπο ώστε να μην μπορεί να εξακριβωθεί πλέον η ταυτότητα του προσώπου στο οποίο αναφέρονται
- Όμως προσοχή: Δεν είναι εύκολος ο χαρακτηρισμός των δεδομένων ως ανώνυμων;
  - Πρέπει να ληφθεί υπόψη κάθε εύλογο μέσο που μπορεί να χρησιμοποιήσει κανείς για να αναγνωρίσει κάποιο πρόσωπο
  - Το ότι νομίζουμε ότι είναι ανώνυμα, ίσως να μην αρκεί..

Ανώνυμα δεδομένα: το συχνό λάθος

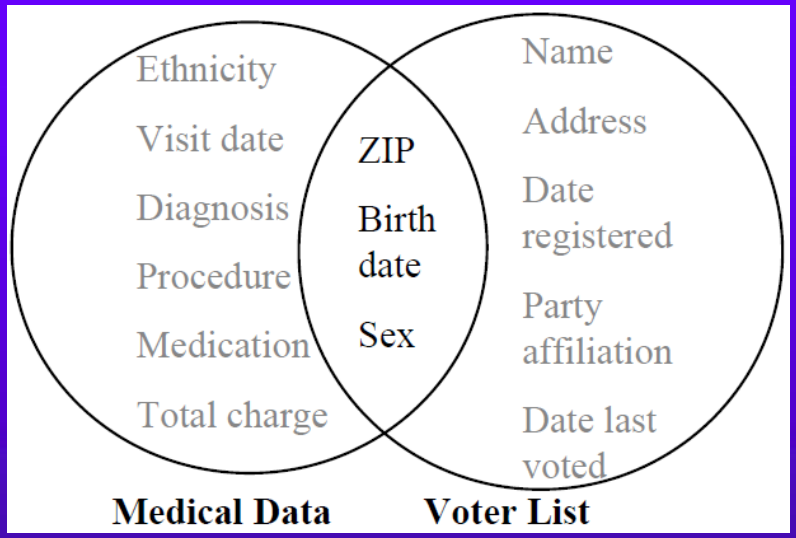
- Το (συχνό) λάθος: Αν δεν είναι «καταφανές» σε ποιον αναφέρονται τα δεδομένα, τότε είναι ανώνυμα

- Το σωστό: Ακόμα και αν δεν είναι προφανής η ταυτότητα του ανθρώπου στον οποίο αναφέρονται τα δεδομένα, πρέπει – προτού χαρακτηριστούν ως ανώνυμα – να εξεταστεί ενδελεχώς αν όντως έχει «εκμηδενιστεί» η δυνατότητα ανακάλυψης της ταυτότητάς του

- Οι «λανθασμένες» ανωνυμοποιήσεις είναι μία ιστορία παλιά….

Επίθεση συσχετίσεων (Linking Attack)

- [Sweeney, 2002] : Αντιπαραβολή της λίστας ψηφοφόρων με την (ανωνυμοποιημένη) λίστα νοσηλευομένων δημόσιου νοσοκομείου



- Για μία συγκεκριμένη ημ/νία γέννησης, έξι άτομα είχαν την ίδια,
  - Τρεις εξ αυτών άντρες, μόνο ένας με τον ίδιο ταχυδρομικό κώδικα (ZIP code)
  - Αυτός ήταν ο (τότε) κυβερνήτης της Μασαχουσέτης
- Σύμφωνα με αυτήν την έρευνα, το 87% του πληθυσμού των Η.Π.Α. μπορεί να ταυτοποιηθεί από την τριπλέτα «Ταχ. Κώδικας - ημερομηνία γέννησης – φύλο»

Το περιστατικό της AOL (2006)

Αύγουστος 2006: research.aol.com

AOL is embarking on a new direction for its business making its content and products freely available to all consumers. To support those goals, AOL is also embracing the vision of an open research community. To get started, we invite you to visit us at http://research.aol.com, where you will find:

- *…*
- ***Query streams for 500,000 users over 3 months (20 million queries)***
- *….*

- Δεν υπήρχαν στη σχετική λίστα των χρηστών άμεσα στοιχεία ταυτοποίησής τους
  - Ένα τυχαίο ID για τον κάθε χρήστη

- Όμως, συνδυάζοντας κανείς τις αναζητήσεις ενός χρήστη (βάσει του ID αυτού) με λοιπές πληροφορίες (π.χ. τηλεφωνικού καταλόγου), μπορούσε τελικά να ταυτοποιήσει το χρήστη αυτόν!!
  - Η εφημερίδα New York Times ταυτοποίησε πολλούς χρήστες
    - Κάποιους εξ αυτών, με την άδειά τους, δημοσιοποίησε τα στοιχεία τους για να καταδείξει το περιστατικό.

Η περίπτωση του χρήστη #4417749

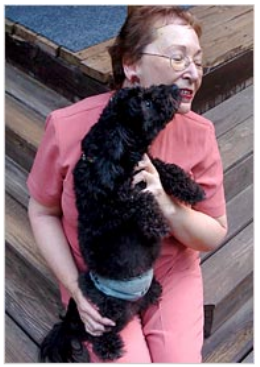
The New York Times

Technology

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Multimedia

Graphic: What Revealing Search Data Reveals

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

AOL removed the search data from its site over the weekend and apologized for its release, saying it was an unauthorized move by a team that had hoped it would benefit academic researchers.

But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans, copies of which continue to circulate online, underscore how much people unintentionally reveal about themselves when they use search engines — and how risky it

http://www.netflixprize.com/

Netflix Prize: Home

NETFLIX

Netflix Prize

Home | Rules | Leaderboard | Register | Update | Submit | Download

Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!

Το περιστατικό της Netflix

- 2006: Η Netflix δημοσιοποίησε τις αξιολογήσεις που έκαναν οι εγγεγραμμένοι σε αυτή χρήστες σε ταινίες
  - Κάθε στοιχείο ταυτοποίησής τους είχε απομακρυνθεί
- Ένα χρόνο μετά (2007), οι ερευνητές Narayanan and Shmatikov ταυτοποίησαν σημαντικό ποσοστό των χρηστών της Netflix, με βάση τις (δημόσια προσβάσιμες) αξιολογήσεις σε ταινίες που έκαναν στην πλατφόρμα IMDB

  "Given a user's public IMDb ratings, which the user posted voluntarily to selectively reveal some of his movie likes and dislikes, we discover all the ratings that he entered privately into the Netflix system, presumably expecting that they will remain private"
- Με την ταυτοποίηση/συσχέτιση, αποκαλύφθηκαν και ευαίσθητα δεδομένα βάσει συγκεκριμένων αξιολογήσεων ταινιών που έγιναν στη Netflix (με την προσδοκία ότι δεν θα δημοσιοποιηθούν).
  - *Π.χ. "Power and Terror: Noam Chomsky in Our Times", "Fahrenheit 9/11", "Jesus of Nazareth"*

Πότε μπορεί να αναγνωριστεί η ταυτότητα ενός προσώπου;

- Εκτός από τα αναγνωριστικά (identifiers), υπάρχουν και τα ψευδο-αναγνωριστικά (quasi-identifiers), που συνδυαστικά δύνανται επίσης να οδηγήσουν σε ταυτοποίηση!

Identifier	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

- Η απαλοιφή των αναγνωριστικών δεν διασφαλίζει εγγυημένα ανωνυμία

Παράδειγμα «κακής ανωνυμοποίησης»

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

- Παράδειγμα: Νοσοκομείο «δημοσιεύει» τον ανωτέρω «ανωνυμοποιημένο» πίνακα (π.χ. τα διαβιβάζει σε ερευνητή για ερευνητικούς σκοπούς)
  - Έχει αφαιρέσει κάθε στοιχείο που θα μπορούσε να οδηγήσει στην ταυτοποίηση (ΑΦΜ, ΑΜΚΑ, Αρ. ταυτότητας, ονοματεπώνυμο)

Πόσο ανώνυμος είναι ο πίνακας;

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Πηγή: B. Fung et.al., Privacy-Preserving Data Publishing: A Survey of Recent Developments, ACM Computing Surveys, 2010

- Έστω ότι ο ερευνητής γνωρίζει ότι στη λίστα του νοσοκομείου υπάρχουν κάποια συγκεκριμένα άτομα (π.χ. οι κάτοικοι ενός μικρού χωριού)
- Για αυτά τα άτομα, μπορεί «εύκολα» (π.χ. από δημόσια προσβάσιμες πηγές) να έχει πρόσβαση σε δεδομένα τους (βλ. Πίνακα b)
  - Συνδυάζοντας τους δύο πίνακες, μπορεί να οδηγηθεί σε ταυτοποίηση κάποιων!!
    - Π.χ. από την τριπλέτα (Job, Sex, Age) = (Laywer, Male, 38) εξάγει το συμπέρασμα για τη νόσο HIV στον Doug

Συσχέτιση εγγραφών

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

- Η περίπτωση αυτή λέγεται συσχέτιση εγγραφών (record linkage)
- Η συσχέτιση αυτή οδηγεί σε αποκάλυψη «ευαίσθητης» πληροφορίας για ένα πρόσωπο (εν προκειμένω, για τον Doug)
  - Στο υπόλοιπο της παρουσίασης, με τον όρο «ευαίσθητο πεδίο» (sensitive attribute) θα καλούμε οποιοδήποτε πεδίο σε έναν πίνακα φέρει κρίσιμη πληροφορία που θέλουμε να «προστατέψουμε» μέσω της ανωνυμοποίησης

Αντιμετώπιση του προβλήματος – «Γενίκευση» γνωρισμάτων

- Για την αποφυγή αυτών των επιθέσεων, μεταβάλλουμε κατάλληλα τις τιμές των πεδίων που είναι quasi-identifiers, μέσω γενίκευσής τους (generalization)
  - Π.χ δεν δημοσιεύουμε επακριβώς την ηλικία, αλλά ένα εύρος ηλικιών (π.χ. 30-40)
  - Με αυτόν τον τρόπο, η μονοσήμαντη συσχέτιση μίας καταχώρησης του «ανώνυμου» πίνακα με μία του «επώνυμου» καθίσταται πιο δύσκολη
    - Όσο πιο μεγάλη η γενίκευση, τόσο ενισχύουμε την ανωνυμοποίηση, αλλά από την άλλη πλευρά έχουμε «απώλεια» χρήσιμης πληροφορίας για ερευνητικούς σκοπούς
    - Στόχος η – κατά το δυνατόν – μέγιστη ανωνυμοποίηση, με τη μικρότερη δυνατή απώλεια πληροφορίας

«Γενίκευση» στον προηγούμενο πίνακα – Δένδρο ταξινόμησης

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

«Γενίκευση»

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

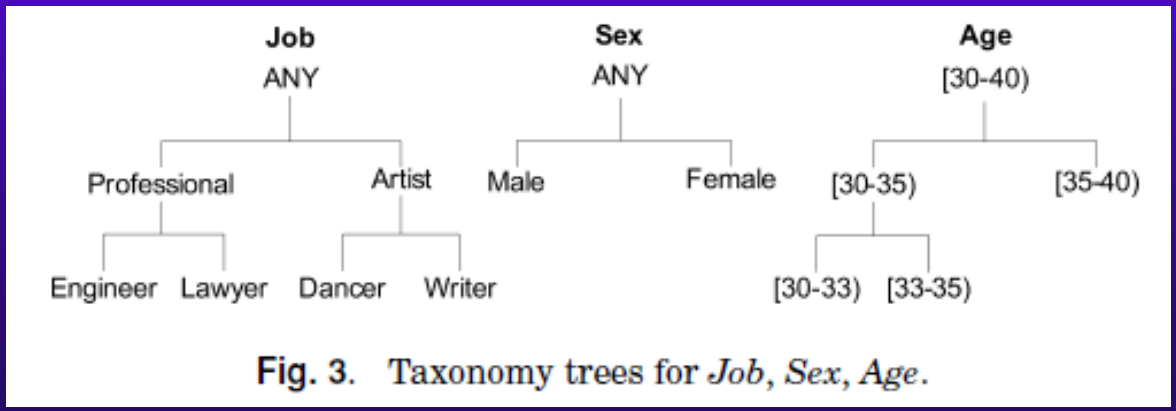


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

Τι κερδίσαμε;

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

??

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

- Η συσχέτιση των δύο πινάκων, συγκρίνοντας τους quasi-identifiers, δεν μπορεί να γίνει

- Στον ανωνυμοποιημένο πίνακα, έχουμε ομάδες εγγραφών, όπου σε κάθε ομάδα υπάρχει ταύτιση στις τιμές των ψευδο-αναγνωριστικών
  - Κάθε τέτοια ομάδα λέγεται κλάση ισοδυναμίας
  - Στον ανωτέρω πίνακα έχουμε δύο κλάσεις ισοδυναμίας

Κριτήρια για τη γενίκευση

- Ανωνυμία k τάξης (k-anonymity) - Samarati-Sweeney, 1998:
  Ικανοποιείται όταν, σε έναν ανώνυμο πίνακα, το σύνολο των εγγραφών (καταχωρήσεων) με τις ίδιες τιμές στα quasi-identifiers είναι τουλάχιστον k
- Προφανώς, όσο μεγαλύτερο το k, τόσο ενισχύεται η ανωνυμία
- Επιστροφή στο προηγούμενο παράδειγμα: Ανώνυμος 3ης τάξης
  - Βλέπετε γιατί;

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

- Στόχος: Ελάχιστη ανωνυμοποίηση k-τάξης (ο πίνακας να μην προκύπτει από γενίκευση ενός άλλου ανώνυμου πίνακα τάξης k)
- Γνωστοί αλγόριθμοι: Incognito, Mondrian

Αντί για γενίκευση;

- Κατάργηση (suppression): Κάποια πεδία (ή και ολόκληρες εγγραφές) αποβάλλονται τελείως

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Flu

Γενίκευση

Κατάργηση πεδίου

- Η γενίκευση στο ανώτατο δυνατό όριο ουσιαστικά ισοδυναμεί με κατάργηση

Είναι αρκετή η ανωνυμοποίηση k τάξης;

- Έστω ότι ξέρουμε τα εξής:

	Zip	Age	National
Bob →	13053	31	American
Akira →	13068	21	Japanese

- Καθώς επίσης και ότι οι Bob και Akira εμπεριέχονται σε έναν πίνακα που θα δημοσιεύσει ο εκδότης (publisher)

Αρχικά δεδομένα

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
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	HIV
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	HIV
10	13053	37	Indian	HIV
11	13068	36	Japanese	HIV
12	13068	35	American	HIV

Έχουν απομακρυνθεί οι identifiers (Οι Bob, Akira έχουν «χρωματιστεί»)

Ανωνυμία 4ης τάξης

	Non-Sensitive Data			Sensitive Data
#	ZIP	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*HIV*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
8	*1485**	*> = 40*	*	*Viral Infection*
9	130**	3*	*	HIV
10	130**	3*	*	HIV
11	130**	3*	*	HIV
12	130**	3*	*	HIV

Η Akira ανήκει εδώ

Ο Bob ανήκει εδώ

“Προστατεύουμε» τα δεδομένα;

	Non-Sensitive Data			Sensitive Data
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*HIV*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
			*	*Viral Infection*
			*	HIV
			*	HIV
11	130**	3*	*	HIV
12	130**	3*	*	HIV

Η Akira ανήκει εδώ

Ο Bob ανήκει εδώ

Ο Bob έχει HIV!!

“Προστατεύουμε» τα δεδομένα;

	Non-Sensitive Data			Sensitive Data
#	ZIP	A	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	> = 40	*	HIV
6	1485*	> = 40	*	Heart Disease
7	1485*	> = 40	*	Viral Infection
8	1485*	> = 40	*	Viral Infection
9	130**	3*	*	HIV
10	130**	3*	*	HIV
11	130**	3*	*	HIV
12	130**	3*	*	HIV

Αν γνωρίζουμε ότι οι παθήσεις της καρδιάς είναι εξαιρετικά σπάνιες στην Ιαπωνία, τότε με πολύ μεγάλη πιθανότητα η Akira έχει μόλυνση από ιό!!!

Η Akira ανήκει εδώ

Ο Bob έχει HIV!!

Ο Bob ανήκει εδώ

Επιθέσεις εξαγωγής συμπεράσματος (Inference attacks)

- Εφαρμόζονται όταν εξάγεται συμπέρασμα για μία «ευαίσθητη» πληροφορία ενός ατόμου, ακόμα και αν δεν αναγνωρίζεται επακριβώς ποια είναι η καταχώρησή του στον ανωνυμοποιημένο πίνακα
- Οι ανωνυμοποιήσεις τάξης k δεν μπορούν να προστατεύσουν ως προς αυτές τις επιθέσεις (βλ. προηγούμενο παράδειγμα)
- Γνωστές επίσης και με το όνομα «attribute linkage»

Ανωνυμία διαφορετικότητας τάξης l (l-diversity)

- Το σύνολο των εγγραφών που είναι «αξεχώριστες» (ίδιες τιμές στα QID) συνιστά μία κλάση ισοδυναμίας
- l-diversity (Machanavajjhala et al., 2006): Κάθε κλάση ισοδυναμίας πρέπει να περιέχει τουλάχιστον l «καλά ορισμένες» διακριτές τιμές του ευαίσθητου πεδίου
  - Πιο απλή περίπτωση: Distinct l-diversity
    - Σε κάθε κλάση ισοδυναμίας εμφανίζονται ακριβώς l διακριτές τιμές από το «ευαίσθητο» πεδίο

Distinct 3-diversity

	Non-Sensitive Data			Sensitive Data
#	ZIP	Age	Nationality	Condition
1	1305*	<= 40	*	Heart Disease
2	1305*	<= 40	*	Viral Infection
3	1305*	<= 40	*	HIV
4	1305*	<= 40	*	HIV
5	*1485**	*>= 40*	*	*HIV*
6	*1485**	*>= 40*	*	*Heart Disease*
7	*1485**	*>= 40*	*	*Viral Infection*
8	*1485**	*>= 40*	*	*Viral Infection*
9	1306*	<= 40	*	Heart Disease
10	1306*	<= 40	*	Viral Infection
11	1306*	<= 40	*	HIV
12	1306*	<= 40	*	HIV

Οι Bob, Akira ανήκουν εδώ

Distinct l-Diversity: Είναι αρκετή;

- Επιθέσεις εξαγωγής συμπεράσματος εξακολουθούν να μπορούν να γίνουν, έστω και πιθανοτικά (probabilistic inference attacks)

10 εγγραφές - Κλάση ισοδυναμίας

...	**Disease**
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

8 από 10 νοσούν από HIV

l-Diversity: Άλλες εκδοχές

- Entropy l-diversity
  - Σε κάθε κλάση ισοδυναμίας, η εντροπία της μεταβλητής που καθορίζεται από το «ευαίσθητο» πεδίο είναι τουλάχιστον ίση με log(l)
- Επιστροφή στο παράδειγμα της 3ης-τάξης ανωνυμίας:

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

  - Πρώτη κλάση ισοδυναμίας:

$$-\frac{2}{3}\log\frac{2}{3}-\frac{1}{3}\log\frac{1}{3}=\log(1.9)$$

  - Δεύτερη κλάση ισοδυναμίας:

$$-\frac{1}{4}\log\frac{1}{4}-\frac{3}{4}\log\frac{3}{4}=\log(1.8)$$

  - Άρα, ο πίνακας είναι “entropy 1.8-diverse” (ενώ είναι 2-diverse)

l-Diversity: Άλλες εκδοχές

- Recursive (*c*,*l*)-diversity:
  - Έστω ότι σε μία κλάση ισοδυναμίας υπάρχουν m διαφορετικές τιμές για το ευαίσθητο πεδίο
  - Έστω *fi το πλήθος φορών της* i-οστής πιο συχνά εμφανιζόμενης τιμής μέσα στην κλάση
  - Μία κλάση ισοδυναμίας είναι (*c*,*l*)-*diverse* εάν η συχνότητα εμφάνισης της πιο συχνής τιμής είναι μικρότερη από το άθροισμα των συχνοτήτων των m-l+1 λιγότερο συχνά εμφανιζόμενων τιμών, δηλαδή: $f1 < c \sum_{i=l}^{m} f_i$
    όπου c μία σταθερά που επιλέγει αυτός που επιτελεί την ανωνυμοποίηση
  - Εάν το παραπάνω ισχύει για κάθε κλάση ισοδυναμίας, τότε όλος ο πίνακας είναι recursive (c,l)-diverse

Επιθέσεις συσχέτισης πινάκων (table linkage)

3-anonymous, 2-diverse

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

Εξωτερικός πίνακας

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

- Η πιθανότητα να εμπεριέχεται η Alice στον ανωνυμοποιημένο πίνακα είναι 4/5
- Πολλές φορές, αναλόγως του είδους των δεδομένων, η πληροφορία αν ένα πρόσωπο εμπεριέχεται ή όχι σε μία ανωνυμοποιημένη βάση μπορεί να είναι εξαιρετικά σημαντική

Περιορισμοί κατά την l-Diversity ανωνυμοποίηση

- Η αρχική κατανομή των τιμών που λαμβάνει το ευαίσθητο πεδίο επιφέρει περιορισμούς
- Σε κάποιες περιπτώσεις, αναλόγως της κρισιμότητας της πληροφορίας και της γενικότερης στατιστικής του συνολικού πίνακα, ενδεχομένως να μη χρειάζεται
- Παράδειγμα: Έστω ότι το ευαίσθητο πεδίο λαμβάνει δύο μόνο διαφορετικές τιμές στον αρχικό πίνακα: HIV+ (1%) και HIV- (99%)
  - Η 2-diversity ανωνυμία είναι περιττή σε μία κλάση ισοδυναμίας που περιέχει μόνο την HIV- στο ευαίσθητο πεδίο
  - Εξάλλου, είναι και δύσκολο να επιτευχθεί:
    - Έστω ότι το σύνολο των καταχωρήσεων είναι 10000
    - Για την επίτευξη distinct 2-diversity, πρέπει να σχηματίσουμε το πολύ 10000*1%=100 κλάσεις ισοδυναμίας

Επιθέσεις ασυμμετρίας (Skewness Attack)

- Παράδειγμα: Δύο πιθανές τιμές για το ευαίσθητο πεδίο: HIV+ (1%) και HIV- (99%)
- Ας θεωρήσουμε μία κλάση ισοδυναμίας που περιέχει ίδιο πλήθος εγγραφών με την τιμή HIV+ όσο και με την τιμή HIV-
  - Αν γνωρίζουμε κάποιος ότι ανήκει στην κλάση, αυτομάτως εξάγουμε το συμπέρασμα ότι έχει ψηλή (σε σχέση με το λοιπό πληθυσμό) πιθανότητα για HIV+
- Προσοχή: Με την l-diversity ανωνυμοποίηση, οι παρακάνω δύο κλάσεις ισοδυναμίας είναι εξίσου αποδεκτές:
  - Κλάση ισοδυναμίας 1: 49 HIV+ and 1 HIV-
  - Κλάση ισοδυναμίας 1 : 1 HIV+ and 49 HIV-

«Αποκάλυψη» ευαίσθητης πληροφορίας

Bob	
Zip	***Age***
47678	27

Distinct 3-diverse πίνακας:
(Δύο «ευαίσθητα» πεδία: μισθός – νόσημα)

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Duodenal Ulcer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Duodenal Ulcer

Τι συμπεράσματα εξάγουμε;

1. Ο Bob ανήκει στους χαμηλόμισθους
2. Ο Bob έχει κάποια αρρώστια που σχετίζεται με το στομάχι

Στην l-diversity δεν λαμβάνεται υπόψη το περιεχόμενο (σημασιολογικά) των τιμών του ευαίσθητου πεδίου

Ανωνυμία t-εγγύτητας (t-closeness)

- (Li et al., 2007) Στόχος: Σε κάθε κλάση ισοδυναμίας, η κατανομή των τιμών του ευαίσθητου πεδίου να προσεγγίζει (ιδανικά, να ταυτίζεται) με την αντίστοιχη κατανομή που εμφανίζεται σε όλον τον πίνακα.
  - Για την αποτροπή επιθέσεων ασυμμετρίας
  - Η «απόσταση» μεταξύ της κατανομής των τιμών του ευαίσθητου πεδίου στην κλάση ισοδυναμίας με αυτή του συνολικού πίνακα δεν ξεπερνά ένα προκαθορισμένο κατώφλι t.

Παράδειγμα t-closeness ανωνυμίας

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

- Προσοχή: Σε αυτήν την περίπτωση, χρειάζονται τα quasi-identifiers;
- Ένας τέτοιος πίνακας δεν μπορεί να αξιοποιηθεί ουσιαστικά για π.χ. ερευνητικούς σκοπούς
  - Θα μπορούσαμε απλά να ανακοινώσουμε την τελευταία στήλη: έχουμε πολύ μεγάλη απώλεια πληροφορίας

Λοιποί κίνδυνοι

Caucas	787XX	Flu	HIV-
Caucas	787XX	Shingles	HIV+
Caucas	787XX	Acne	HIV-
Caucas	787XX	Flu	HIV-
Caucas	787XX	Acne	HIV-
Caucas	787XX	Flu	HIV-
Asian/AfrAm	78XXX	Flu	HIV-
Asian/AfrAm	78XXX	Flu	HIV-
Asian/AfrAm	78XXX	Acne	HIV+
Asian/AfrAm	78XXX	Shingles	HIV-
Asian/AfrAm	78XXX	Acne	HIV-
Asian/AfrAm	78XXX	Flu	HIV-

Bob is Caucasian and
I heard he was admitted to hospital
with Shingles…

Το πρόβλημα οφείλεται στο ότι δύο πεδία δεν είναι quasi-identifiers

Συμπεράσματα μέχρι τώρα

- Δεν υπάρχει συγκεκριμένη κατηγορία ανωνυμοποίησης που να επιτυγχάνει ένα βέλτιστο αποτέλεσμα
- Αναλόγως του είδους των δεδομένων και την εφαρμογή στην οποία αναφέρονται (σκοπός επεξεργασίας), πρέπει κάθε φορά να επιλέγεται η κατάλληλη
  - Προσοχή χρειάζεται στη διαδικασία γενίκευσης ή/και κατάργησης (λαμβάνοντας πάντα υπόψη και το ανεκτό επίπεδο απώλειας πληροφορίας)
- Τα κύρια χαρακτηριστικά των κατηγοριών ανωνυμοποίησης, καθώς και οι κίνδυνοι ως προς την ιδιωτικότητα που ελοχεύουν από τον καθένα, πρέπει να είναι εις γνώσιν του «εκδότη» (data publisher)

Απώλεια πληροφορίας

- Κάθε γενίκευση γνωρισμάτων προκαλεί κάποια απώλεια χρήσιμης πληροφορίας (information loss)
  - Πώς αυτή αποτιμάται;
- Υπάρχουν διάφορες μετρικές για την απώλεια της πληροφορίας
- Normalized certainty penalty (NCP) – Κανονικοποιημένη ποινή βεβαιότητας
  - Συνάρτηση, κατά κάποιο τρόπο, του «βάθους» του δέντρου γενίκευσης
  - Εφαρμόζεται τόσο σε γενικεύσεις αριθμητικών τιμών όσο και σε γενικεύσεις μη αριθμητικών τιμών για τα γνωρίσματα
- Όσο μικρότερη τιμή λαμβάνει, τόσο μικρότερη απώλεια πληροφορίας έχουμε

NCP για αριθμητικές τιμές

- Έστω πίνακας T με τα αριθμητικά ψευδο-αναγνωριστικά (A1…….An).
- Έστω μία εγγραφή t = (x1…….xn) «γενικεύεται» σε μία νέα εγγραφή t′ = ([y1,z1],……[yn,zn]), όπου yi ≤ xi ≤ zi (1≤ i≤n).
- Ο δείκτης NCP της εγγραφής t για το γνώρισμα Ai ισούται με

$$\mathrm{NCP}_{Ai(t)} = \frac{z_i - y_i}{|Ai|}, \text{ όπου } |\mathrm{Ai}| = \max_{t \in T} t.Ai - \min_{t \in T} t.Ai$$

- Ουσιαστικά, σε μία κλάση ισοδυναμίας, όλες οι εγγραφές έχουν ακριβώς την ίδια τιμή NCP για το συγκεκριμένο γνώρισμα Ai

NCP για αριθμητικές τιμές - Παράδειγμα

$$\text{NCP}_{Ai(t)} = \frac{z_i - y_i}{|Ai|}, \text{ όπου } |\text{Ai}| = \max_{t \in T} t.Ai - \min_{t \in T} t.Ai$$

- $\text{NCP}_{\text{Age(t)}} = 5/8$ για κάθε εγγραφή t
  - Κάθε διάστημα γενίκευσης έχει τιμή 5 ([35-40], [30-35])
  - Η μικρότερη ηλικία του πίνακα είναι 30 και η μεγαλύτερη 38 (οπότε και 38-30=8)

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

«Γενίκευση»

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

NCP για μη αριθμητικές τιμές

- Έστω πίνακας T με μη αριθμητικό ψευδο-αναγνωριστικό Ai.
- Έστω μία εγγραφή t που έχει αρχικώς τιμή v στο Ai, αλλά με τη γενίκευση η τιμή που λαμβάνει είναι μία εκ των (v1……vm)
- Στο δέντρο γενίκευσης, βρίσκουμε το πλήθος των αρχικών τιμών του πίνακα τα οποία «καλύπτονται» από τη γενίκευση (v1……vm)
  - Συμβολίζεται με (| ancestor(v1……vm) |)
- Ο δείκτης NCP της εγγραφής t για το γνώρισμα Ai ισούται με

$$\mathrm{NCP}_{Ai(t)} = \frac{|\ \text{ancestor}(v_1 \ldots\ldots v_m)\ | - 1}{|Ai|}$$

, όπου |Ai| το αρχικό πλήθος των τιμών

- Ομοίως με προηγούμενα, η τιμή αυτή είναι ίδια, για δοθέν γνώρισμα, για κάθε εγγραφή της κλάσης ισοδυναμίας

NCP για μη αριθμητικές τιμές

- $NCP_{Επάγγελμα(t)} = 2/4 = 0.5$ για κάθε εγγραφή t
  - Τόσο η τιμή «Artist», όσο και η τιμή «Professional», καλύπτουν ακριβώς δύο από τις αρχικές τιμές
  - Όλες οι πιθανές τιμές του επαγγέλματος αρχικώς είναι 4
- Στην κορυφή του δέντρου («ANY») έχουμε τη μέγιστη δυνατή τιμή NCP=1

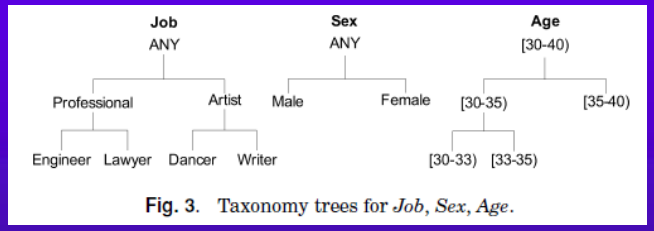


Fig. 3. Taxonomy trees for *Job*, *Sex*, *Age*.

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

«Γενίκευση»

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

οίηση προσωπικώ

Συνολική NCP για κλάση ισοδυναμίας

- Μία εγγραφή έχει διαφορετικές NCP για κάθε γνώρισμα
- Η συνολική NCP μίας εγγραφής t προκύπτει από τη σχέση
- $NCP(t) = \sum_{i=1}^{n} w_i NCP_{Ai}(t)$
- Όπου wi το «βάρος» (weight) κάθε γνωρίσματος, αναλόγως της σημασίας του
  - Για κάποια γνωρίσματα ενδεχομένως να είμαστε πιο «ελαστικοί» ως προς την απώλεια πληροφορίας σε σχέση με κάποια άλλα
- Οι συντελεστές βαρύτητας πρέπει να ικανοποιούν τη σχέση $\sum_{i=1}^{n} w_i = 1$
- Προφανώς, η τιμή αυτή είναι ενιαία για κάθε εγγραφή μίας κλάσης ισοδυναμίας

Συνολική «ποινή» βεβαιότητας για τον πίνακα

- Κάθε κλάση ισοδυναμίας G έχει τη δική της τιμή NCP
  - Συμβολίζεται με *NCP(G)*
- Η συνολική ποινή βεβαιότητας (Global Certainty Property – GCP) για όλον τον πίνακα *P* δίνεται από τη σχέση:

$$GCP(\mathcal{P}) = \frac{\sum_{G \in \mathcal{P}} |G| \cdot NCP(G)}{d \cdot N},$$

- d: το πλήθος των ψευδο-αναγνωριστικών
- N: το πλήθος των εγγραφών στον αρχικό πίνακα
- |G|: η πληθικότητα της κλάσης ισοδυναμίας G

Το GPP λαμβάνει τιμές από 0 (καμία γενίκευση) μέχρι 1 (πλήρεις γενικεύσεις στην τιμή ANY)

Άλλες μέθοδοι εκτός από γενίκευση/κατάργηση;

- Προσθήκη θορύβου: Κάποιες τιμές γνωρισμάτων τροποποιούνται εσκεμμένα, έτσι ώστε να επιτευχθεί ανωνυμοποίηση (π.χ. κατάλληλες κλάσεις ισοδυναμίας που να οδηγούν σε l-diversity ανωνυμοποίηση), χωρίς να μεταβάλλονται ουσιαστικά τα δεδομένα
  - Π.χ. μεταβάλλουμε το πεδίο του ύψους ενός ατόμου κατά 5-6 εκ.

- Ανατομία (Anatomy – Xiao,Tao, 2006): Στόχος είναι η απόκρυψη των συσχετίσεων μεταξύ των εγγραφών μίας κλάσης ισοδυναμίας και των τιμών του ευαίσθητου πεδίου

Περιγραφή της ανατομίας

Αρχικός πίνακας

Α/Α	Ηλικία	Ταχυδρ. Κώδικας	Φύλο	Μισθός
1	25	14540	Θήλυ	500
2	27	14530	Άρρεν	1000
3	34	14550	Άρρεν	1000
4	31	14544	Άρρεν	800
5	37	17430	Θήλυ	950
6	39	18600	Θήλυ	900
7	40	17650	Άρρεν	900
8	38	18200	Θήλυ	700

4-anonymous, 3-diverse

Α/Α	Ηλικία	Ταχυδρ. Κώδικας	Φύλο	Μισθός
1	<=35	[14530-14550]	*	500
2	<=35	[14530-14550]	*	1000
3	<=35	[14530-14550]	*	1000
4	<=35	[14530-14550]	*	800
5	>35	[17430-18600]	*	950
6	>35	[17430-18600]	*	900
7	>35	[17430-18600]	*	900
8	>35	[17430-18600]	*	700

- Με βάση τις κλάσεις ισοδυναμίας του ανώνυμου πίνακα (που προέκυψε μέσω γενίκευσης), κατασκευάζουμε και δημοσιεύουμε τους ακόλουθους πίνακες:

Πίνακας ψευδο-αναγνωριστικών

Α/Α	Ηλικία	Ταχυδρ. Κώδικας	Φύλο	Αριθμός Κλάσης Ισοδυναμίας
1	25	14540	Θήλυ	1
2	27	14530	Άρρεν	1
3	34	14550	Άρρεν	1
4	31	14544	Άρρεν	1
5	37	17430	Θήλυ	2
6	39	18600	Θήλυ	2
7	40	17650	Άρρεν	2
8	38	18200	Θήλυ	2

Πίνακας ευαίσθητων τιμών

Ομάδα	Μισθός	Αριθμός Εμφανίσεων
1	500	1
1	1000	2
1	800	1
2	950	1
2	900	2
2	700	1

Τι «κερδίζουμε» με την ανατομία

Πίνακας ψευδο-αναγνωριστικών

A/A	Ηλικία	Ταχυδρ. Κώδικας	Φύλο	Αριθμός Κλάσης Ισοδυναμίας
1	25	14540	Θήλυ	1
2	27	14530	Άρρεν	1
3	34	14550	Άρρεν	1
4	31	14544	Άρρεν	1
5	37	17430	Θήλυ	2
6	39	18600	Θήλυ	2
7	40	17650	Άρρεν	2
8	38	18200	Θήλυ	2

Πίνακας ευαίσθητων τιμών

Ομάδα	Μισθός	Αριθμός Εμφανίσεων
1	500	1
1	1000	2
1	800	1
2	950	1
2	900	2
2	700	1

- Δεν προσφέρει καλύτερη ανωνυμία από τον 4-anonymous, 3-diverse πίνακα
- Όμως, αφού δεν γίνεται γενίκευση, έχουμε μικρότερη απώλεια πληροφορίας και, συνεπώς, πιο «χρήσιμα» δεδομένα για π.χ. ερευνητικούς ή στατιστικούς σκοπούς

4-anonymous, 3-diverse

A/A	Ηλικία	Ταχυδρ. Κώδικας	Φύλο	Μισθός
1	<=35	[14530-14550]	*	500
2	<=35	[14530-14550]	*	1000
3	<=35	[14530-14550]	*	1000
4	<=35	[14530-14550]	*	800
5	>35	[17430-18600]	*	950
6	>35	[17430-18600]	*	900
7	>35	[17430-18600]	*	900
8	>35	[17430-18600]	*	700

Τι «χάνουμε» με την ανατομία

tuple ID	Age	Sex	Zipcode	Disease
1	[21, 60]	M	[10001, 60000]	pneumonia
2	[21, 60]	M	[10001, 60000]	dyspepsia
3	[21, 60]	M	[10001, 60000]	dyspepsia
4	[21, 60]	M	[10001, 60000]	pneumonia
5	[61, 70]	F	[10001, 60000]	flu
6	[61, 70]	F	[10001, 60000]	gastritis
7	[61, 70]	F	[10001, 60000]	flu
8	[61, 70]	F	[10001, 60000]	bronchitis

Table 2: A 2-diverse table

row #	Age	Sex	Zipcode	Group-ID
1	23	M	11000	1
2	27	M	13000	1
3	35	M	59000	1
4	59	M	12000	1
5	61	F	54000	2
6	65	F	25000	2
7	65	F	25000	2
8	70	F	30000	2

(a) The quasi-identifier table (QIT)

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

(b) The sensitive table (ST)

P(Alice is in the table) = 4/5

P(Alice is in the table) = 1

Name	Age	Sex	Zipcode
Ada	61	F	54000
Alice	65	F	25000
Bella	65	F	25000
Emily	*67*	*F*	*33000*
Stephanie	70	F	30000
...	...	...	...

Table 5: The voter registration list (publicly accessible)

Διαφορική ιδιωτικότητα

- Ένα διαφορετικό μοντέλο: Διαφορική ιδιωτικότητα (differential privacy – Dwork, 2006).
- Ο «εκδότης» δεν ανωνυμοποιεί τη βάση δεδομένων
  - Την τηρεί αυτούσια, αλλά δεν τη δημοσιοποιεί
- Οι τρίτοι («αποδέκτες») μπορούν να υποβάλλουν ερωτήματα στατιστικής φύσης προς τον «εκδότη»
  - Π.χ. πόσα άτομα της βάσης νοσούν από έλκος;
- Αν γίνουν πολλές (κατάλληλες) ερωτήσεις, από το συνδυασμό των απαντήσεων μπορεί να υπάρξει παραβίαση της ιδιωτικότητας για κάποιο πρόσωπο!!
  - Π.χ Ερ. 1: πόσο είναι το άθροισμα των μισθών των υπαλλήλων της εταιρείας;
  - Ερ.2 : πόσο είναι το άθροισμα των μισθών των υπαλλήλων της εταιρείας, μετά την αποχώρηση του Mr. Brown;
- Η διαφορική ιδιωτικότητα αποσκοπεί στη διαφύλαξη της ανωνυμίας σε αυτό ακριβώς το μοντέλο
  - Πρέπει να «προσεχθούν» οι απαντήσεις

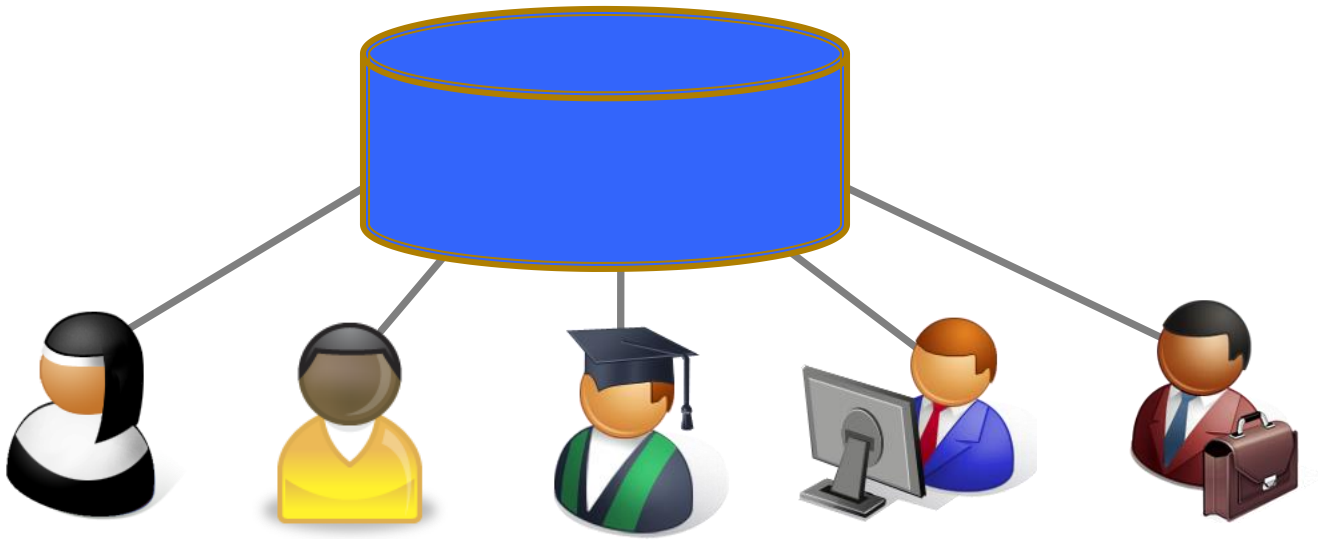
Roadmap

- Defining Differential Privacy
- Techniques for Achieving DP
  - Output perturbation
  - Input perturbation
  - Perturbation of intermediate values
  - Sample and aggregate

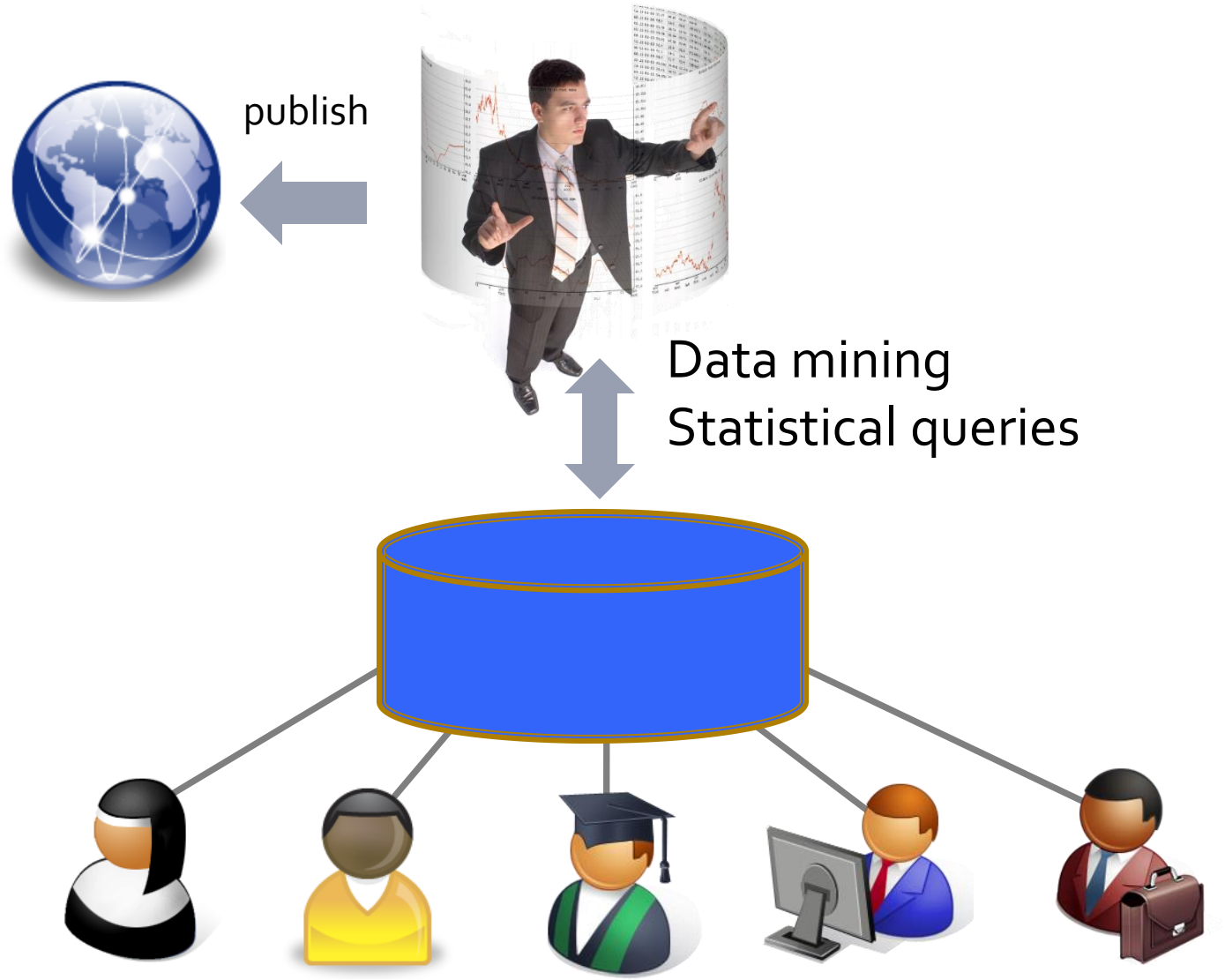
General Setting

Medical data
Query logs
Social network data
…

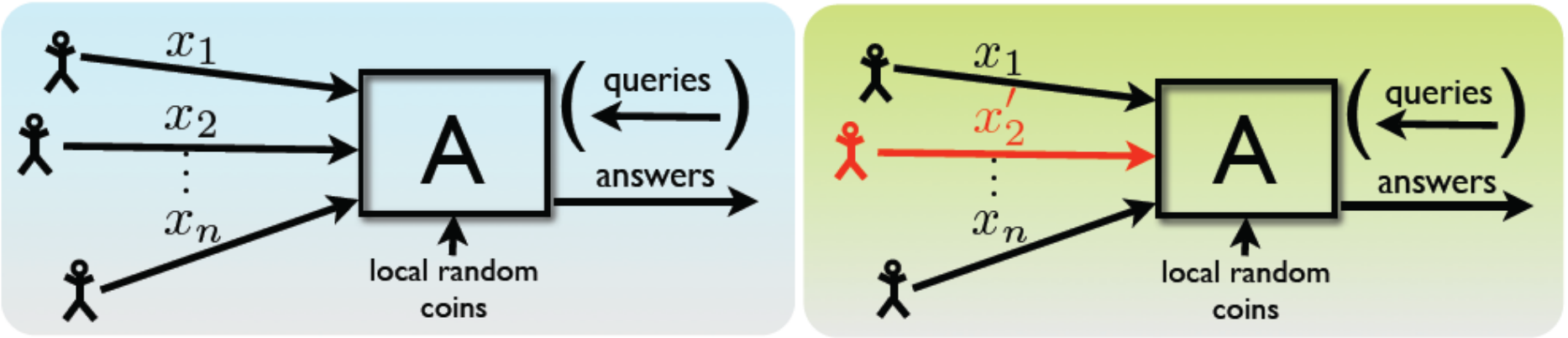
Data mining
Statistical queries



General Setting



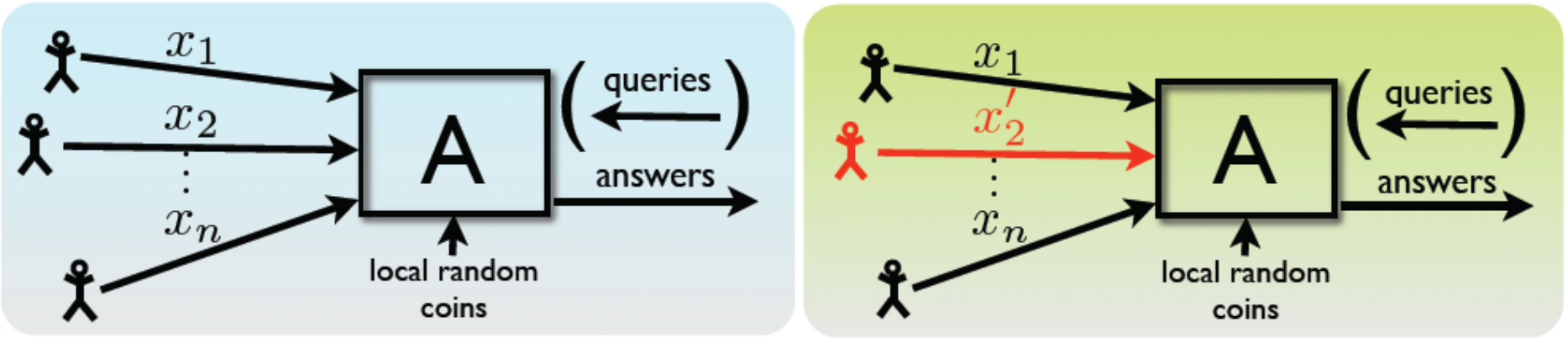
Differential Privacy



x' is a neighbor of x
if they differ in one row

From the released statistics, it is hard to tell which case it is.

Differential Privacy

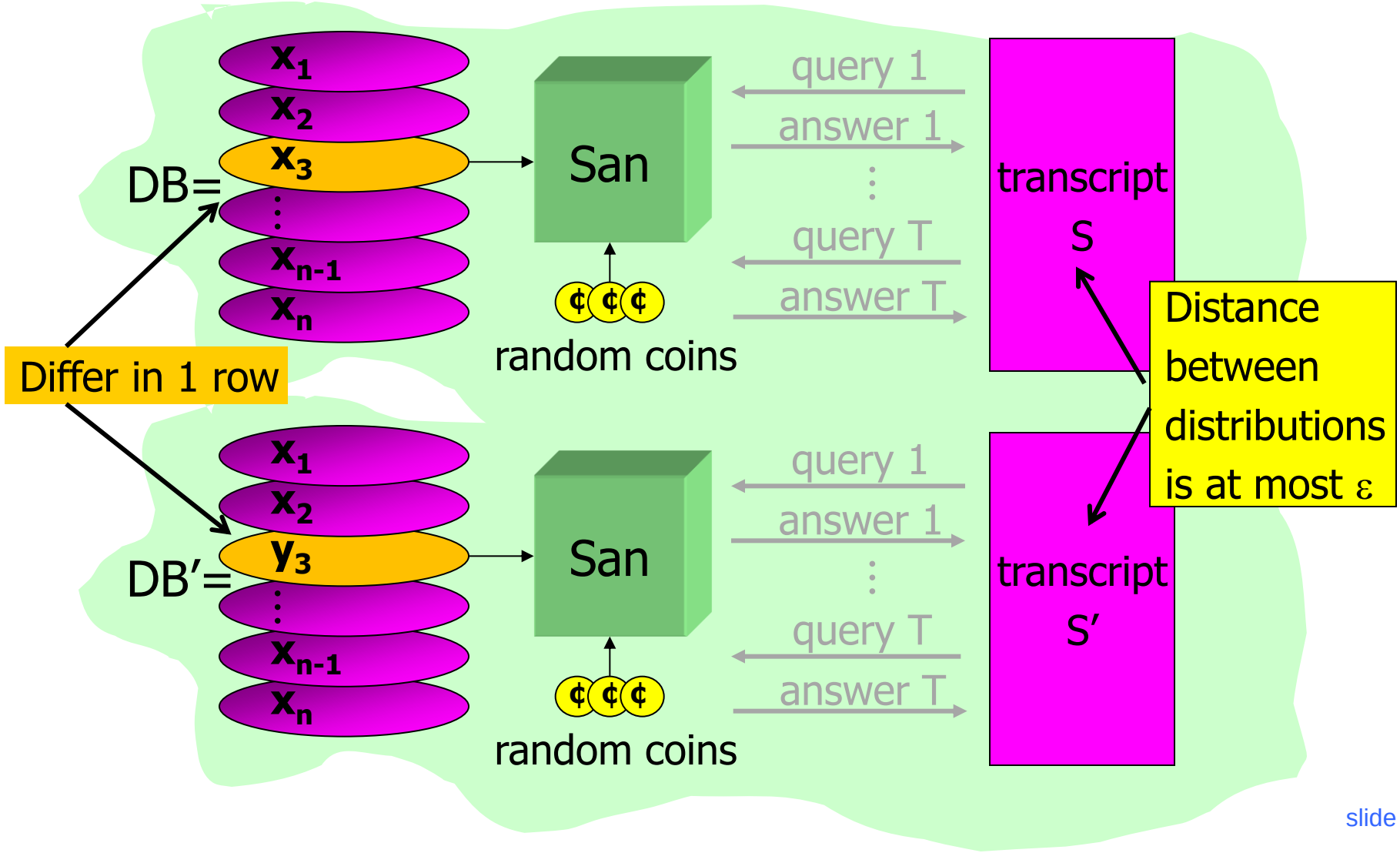


x' is a neighbor of x
if they differ in one row

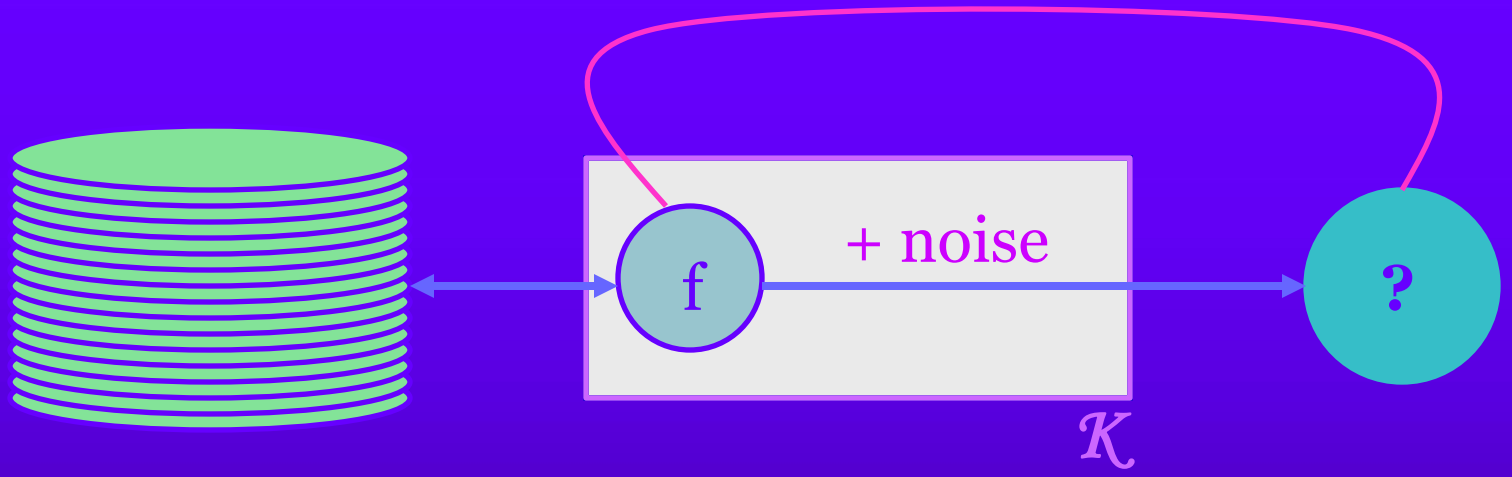
For all neighboring databases x and x'
For all subsets of transcripts:

$$\Pr[A(x) \in S] \leq e^{\epsilon} \Pr[A(x') \in S]$$

Indistinguishability



Προσθήκη θορύβου



An Interactive Sanitizer: K

Dwork, McSherry, Nissim, Smith - 2006

ε-differential privacy:

$$\frac{\Pr[\mathcal{K}(\text{DB} - \text{Me}) = t]}{\Pr[\mathcal{K}(\text{DB} + \text{Me}) = t]} \approx 1 \pm \varepsilon$$

Συμπεράσματα

- Η ανωνυμοποίηση είναι μία εξαιρετικά δύσκολη υπόθεση
- Πρέπει να λαμβάνονται με πάρα πολύ προσοχή τα κατάλληλα μέτρα
  - Και πάλι, πρέπει να έχουμε κατά νου ότι, πιθανότατα, το καλύτερο που μπορούμε να επιτύχουμε είναι να περιορίσουμε τους κινδύνους και όχι να τους εξαλείψουμε
- Εξαιρετικά σημαντικό σε σχέση με τη νομοθεσία προσωπικών δεδομένων
  - Αν δεν είναι ανώνυμα, εφαρμόζεται η νομοθεσία
    - Η οποία με τη σειρά τους επιβάλλει περιορισμούς! Π.χ. είναι νόμιμη η δημοσιοποίηση μη ανώνυμων δεδομένων;

https://arx.deidentifier.org/

ARX

Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface. You can find further information here, or directly proceed to our downloads section.

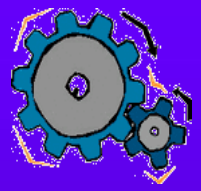
Ψευδωνυμοποιημένα δεδομένα

- Στον GDPR εμφανίζεται ο ορισμός της ψευδωνυμοποίησης:
  - *η επεξεργασία δεδομένων προσωπικού χαρακτήρα κατά τρόπο ώστε τα δεδομένα να μην μπορούν πλέον να αποδοθούν σε συγκεκριμένο υποκείμενο των δεδομένων χωρίς τη χρήση συμπληρωματικών πληροφοριών, εφόσον οι εν λόγω συμπληρωματικές πληροφορίες διατηρούνται χωριστά και υπόκεινται σε τεχνικά και οργανωτικά μέτρα προκειμένου να διασφαλιστεί ότι δεν μπορούν να αποδοθούν σε ταυτοποιημένο ή ταυτοποιήσιμο φυσικό πρόσωπο*
- Στον GDPR αναφέρεται ρητώς ότι τα ψευδωνυμοποιημένα δεδομένα δεν πρέπει να θεωρούνται ανώνυμα
  - *Τα δεδομένα προσωπικού χαρακτήρα που έχουν υποστεί ψευδωνυμοποίηση, η οποία θα μπορούσε να αποδοθεί σε φυσικό πρόσωπο με τη χρήση συμπληρωματικών πληροφοριών, θα πρέπει να θεωρούνται πληροφορίες σχετικά με ταυτοποιήσιμο φυσικό πρόσωπο.*

Ψευδωνυμοποίηση – Ένα απλό σενάριο

Initial data

Mary Adams	Female	23
John Brown	Male	26
Anna Frank	Female	32
Tom Hill	Male	42
. . . .		
. . . .		



Pseudonymous data

A	Female	23
B	Male	26
C	Female	32
D	Male	42
. . . .		
. . . .		

- Οι αντιστοιχίσεις των αναγνωριστικών (Mary Adams, John Brown, …) με τα ψευδώνυμά τους (A, B, …) πρέπει να είναι, κατά κάποιο τρόπο, "προστατευμένες"
  - Τα ψευδώνυμα μπορούν ενδεχομένως να αντικαθιστούν περισσότερα από ένα αναγνωριστικά ή/και ψευδο-αναγνωριστικά
  - Προσοχή: Για μία επιτυχή ψευδωνυμοποίηση, ενδεχομένως να πρέπει να χρησιμοποιηθεί και κάποια τεχνική ανωνυμοποίησης

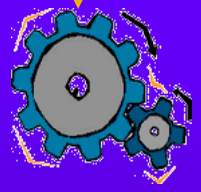
Ψευδωνυμοποίηση ≠ Κρυπτογράφηση

Τυπικό κρυπτογραφικό σχήμα

Initial data

Mary Adams Female 23
John Brown Male 26
Anna Frank Female 32
Tom Hill Male 42
. . . .
. . . .

Encryption key



Encrypted data

hIwDY32hYGCE8MkB
A/wOu7d45aUxF4Q0R
KJprD3v5Z9…

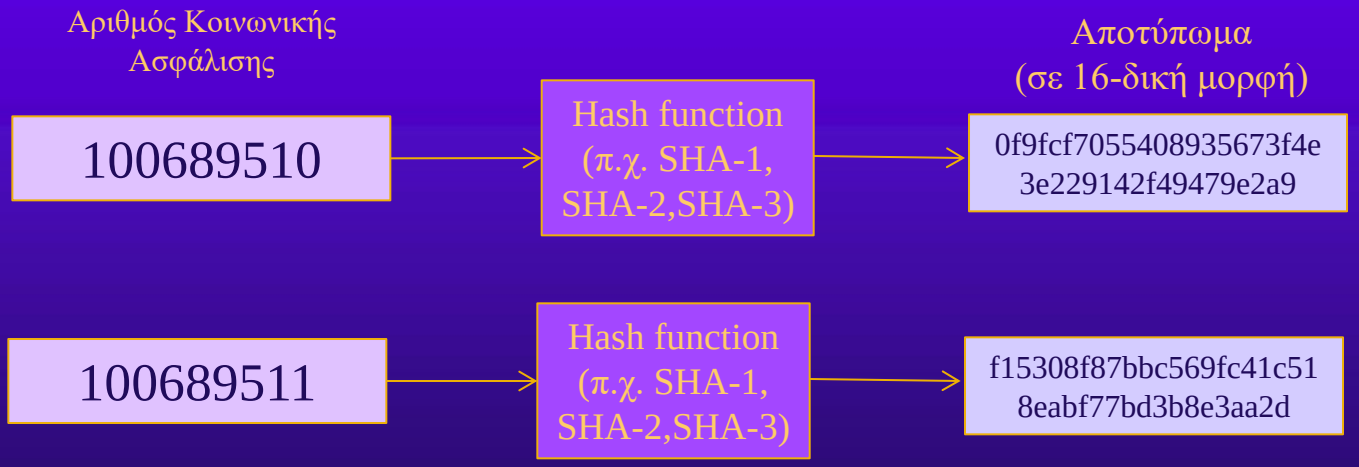
- Η κρυπτογράφηση καθιστά ολόκληρα τα δεδομένα «μη αναγνώσιμα»
- Ουσιαστικά δεν μπορεί να γίνει κάποια στατιστική ανάλυση επί των κρυπτογραφημένων δεδομένων – πρέπει υποχρεωτικά να αποκρυπτογραφηθούν πρώτα (απαιτείται το κλειδί αποκρυπτογράφησης)
  - Αν και υπάρχουν κρυπτογραφικές τεχνικές που «επιτρέπουν» κάποιες πράξεις επί κρυπτογραφημένων μηνυμάτων (π.χ. ομομορφική κρυπτογραφία), εν τούτοις η διαφορά ψευδωνυμοποιημένων με κρυπτογραφημένων δεδομένων είναι καταφανής

Συναρτήσεις κατακερματισμού ως τεχνικές ψευδωνυμοποίησης

- Πολλοί θεωρούν ότι μία κρυπτογραφική συνάρτηση κατακερματισμού, επειδή μαθηματικά είναι μη αντιστρεπτή, μπορεί να χρησιμοποιηθεί για ψευδωνυμοποίηση προσωπικών δεδομένων
  - Π.χ. το αποτύπωμα ενός αριθμού ταυτότητας είναι μη αντιστρέψιμο
- Έχουν το πλεονέκτημα ότι θα αποδίδεται πάντα το ίδιο ψευδώνυμο στον ίδιο χρήστη – σε ορισμένες περιπτώσεις, αυτό είναι αναγκαίο
- Αν όμως είναι κρίσιμο το να μην είναι εφικτή η άρση της ψευδωνυμοποίησης από τρίτους, είναι καλή μέθοδος;

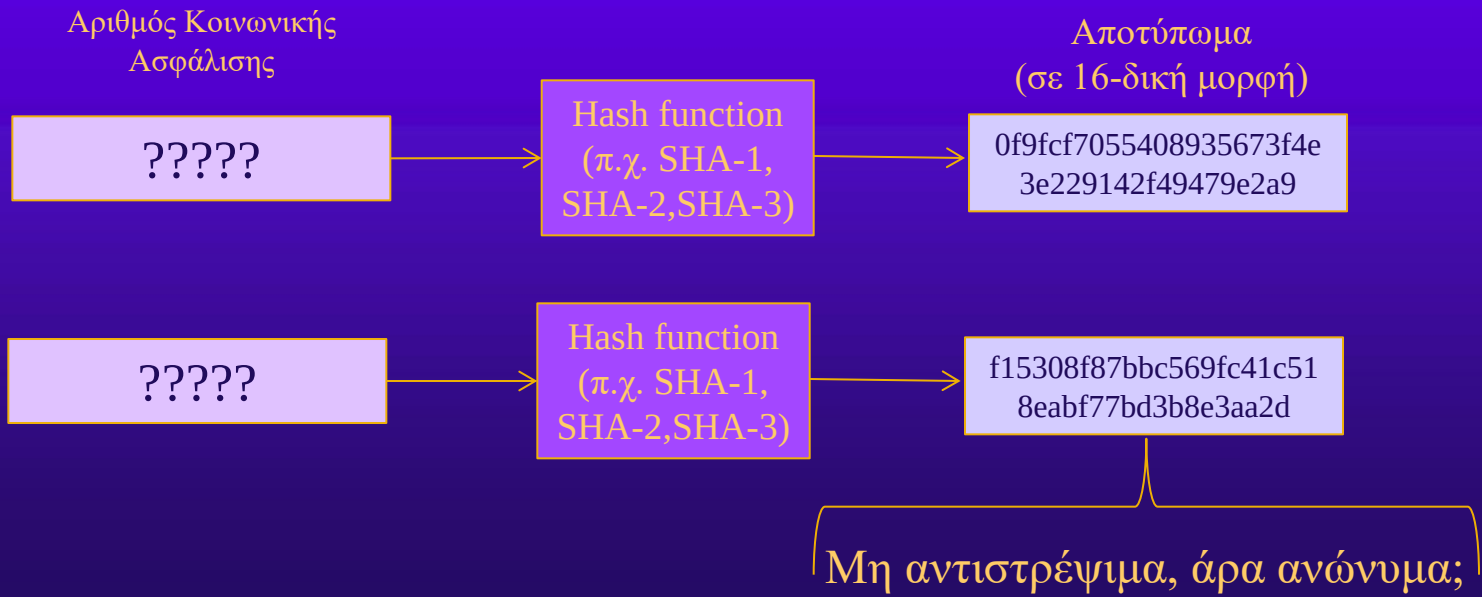
Συναρτήσεις κατακερματισμού ως τεχνικές ψευδωνυμοποίησης

Παράδειγμα: Ερευνητές θέλουν να συλλέξουν (και ενδεχομένως να δημοσιεύσουν) ψευδωνυμοποιημένα δεδομένα. Θέλουν να αποδώσουν ένα μοναδικό ακατάληπτο ψευδώνυμο σε κάθε χρήστη που συμμετείχε στην έρευνα. Θεωρούν ότι μία κρυπτογραφική συνάρτηση κατακερματισμού είναι καλή επιλογή, διότι είναι μη αναστρέψιμη και ταυτόχρονα εξασφαλίζει ότι το ίδιο ψευδώνυμο θα αποδίδεται πάντα στο ίδιο πρόσωπο



Συναρτήσεις κατακερματισμού ως τεχνικές ψευδωνυμοποίησης

Παράδειγμα: Ερευνητές θέλουν να συλλέξουν (και ενδεχομένως να δημοσιεύσουν) ψευδωνυμοποιημένα δεδομένα. Θέλουν να αποδώσουν ένα μοναδικό ακατάληπτο ψευδώνυμο σε κάθε χρήστη που συμμετείχε στην έρευνα. Θεωρούν ότι μία κρυπτογραφική συνάρτηση κατακερματισμού είναι καλή επιλογή, διότι είναι μη αναστρέψιμη και ταυτόχρονα εξασφαλίζει ότι το ίδιο ψευδώνυμο θα αποδίδεται πάντα στο ίδιο πρόσωπο



Συναρτήσεις κατακερματισμού ως τεχνικές ψευδωνυμοποίησης

Μπορούμε να διαπιστώσουμε αν ο Α με Αριθμό Κοινωνικής Ασφάλισης (Α.Κ.Α.) 100689511 βρίσκεται στη λίστα;

1. 1) Υπολογίζουμε το αποτύπωμα του Α.Κ.Α. του χρήστη Α

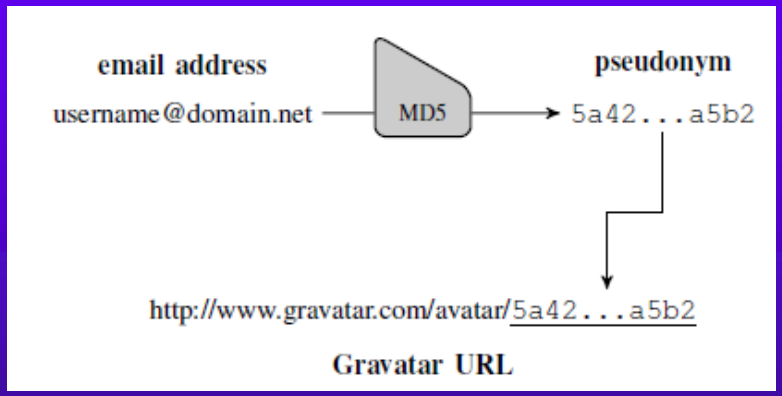
- 2) Ελέγχουμε αν το αποτύπωμα είναι στην «ανωνυμοποιημένη» λίστα

Άρα, αυτή η καταχώρηση αναφέρεται στο χρήστη Α!

Η περίπτωση του Gravatar

- Το Gravatar είναι μία υπηρεσία που επιτρέπει σε χρήστες blogs, forums να έχουν αυτόματα την ίδια εικόνα (“avatar”) για το προφίλ τους, αρκεί να χρησιμοποιούν πάντα την ίδια ηλεκτρονική διεύθυνση στα blogs/forums.
- Αυτές οι εικόνες για τον κάθε χρήστη είναι δημόσια προσβάσιμες με το εξής URL: https://www.gravatar.com/digest

όπου ως «digest» είναι το αποτύπωμα (με MD5) της ηλεκτρονικής διεύθυνσης του χρήστη.



Source: Demir et. al., 2018

- Ερευνητές κατάφεραν να υπολογίσουν τις ηλεκτρονικές διευθύνσεις χρηστών από τα «αποτυπώματα» αυτά – δηλαδή αντιστοίχισαν e-mails με gravatars (εικόνες), χωρίς να έχουν εγγραφεί στα σχετικά blog/forum
- Πιθανή λύση: Hash function με κλειδί

L. Demir, A. Kumar, M. Cunche, C. Lauradoux, “The Pitfalls of Hashing for Privacy”, *Communications Surveys and Tutorials, IEEE Communications Society*, Institute of Electrical and Electronics Engineers, 2018.

Συμπεράσματα

- Η ψευδωνυμοποίηση αποτελεί πολύ σημαντικό εργαλείο ενίσχυσης της ιδιωτικότητας
  - Δεν αποτελεί όμως ανωνυμοποίηση!
  - Ενδεχομένως να πρέπει να συνδυαστεί, αναλόγως των κινδύνων από την εκάστοτε επεξεργασία, και με τεχνικές ανωνυμοποίησης
- Αν και προφανώς δεν είναι κατά κανόνα υποχρεωτική, εν τούτοις είναι πιθανό να αποτελεί κάποιες φορές αναγκαία προϋπόθεση για τη νομιμότητα της επεξεργασίας
  - Π.χ. αναγκαία η ψευδωνυμοποίηση δεδομένων κατά την επεξεργασία δεδομένων για ερευνητικούς σκοπούς
- Σχεδιαστική πρόκληση η εύρεση, από τη στιγμή που θα κριθεί ότι πρέπει να γίνει ψευδωνυμοποίηση, της βέλτιστης τεχνικής ψευδωνυμοποίησης,

Αναφορές

1. Li, N., Li, T., and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 2007.
2. Machanavajhala, A., Kifer, D, Gehrke, J., and Venkitasubramaniam, M., l-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), 2006.
3. Samarati, P. Protecting Respondents' Identities in Microdata Release. IEEE TKDE, 13(6):1010-1027, 2001.
4. Sweeney, L. *k*-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.
5. Sweeney, L. *k*-Anonymity: Achieving *k*-Anonymity Privacy Protection using Generalization and Suppresion. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.
6. S. G. Tsiafoulis and V. C. Zorkadis. A neural network clustering based algorithm for privacy preserving data mining. Int. Conf. on Computational Intelligence and Security (CIS), 2010.
7. Xiao, X, Tao, Y. Anatomy: Simple and Effective Privacy Preservation. VLDB, 2006.
8. Xiao, X, Tao, Y. *m*-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets. SIGMOD, 2007.