Ανωνυμοποίηση και ψευδωνυμοποίηση δεδομένων προσωπικού χαρακτήρα

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

Δρ. Κωνσταντίνος Λιμνιώτης
Ε.Ε.Π., Πληροφορικός
klimniotis at dpa.gr

eDemocracy
7th International Conference on eDemocracy
Privacy-Preserving, Secure, Intelligent eGovernment Services
Workshop: General Data Protection Regulation

Επισκόπηση παρουσίασης

- Εισαγωγή - Βασικές έννοιες
  - Παραδείγματα ανεπαρκούς ανωνυμοποίησης
  - Η εποχή των «μεγάλων δεδομένων» (Big Data)
  - Νέος Κανονισμός (ΕΕ) 679/2016 (GDPR)
  - Προσωπικά δεδομένα – Ανώνυμα δεδομένα
  - Ψευδωνυμοποιημένα δεδομένα
  - Τεχνικές ανωνυμοποίησης / ψευδωνυμοποίησης
- Διαχείριση αναγνωριστικών
- Διαχείριση ψευδο-αναγνωριστικών
- Νέες τεχνολογίες
  - Γράφοι κοινωνικής δικτύωσης
  - Τεχνολογίες blockchain
- Συμπεράσματα

Εισαγωγικά στοιχεία – Η σημασία της ανωνυμοποίησης

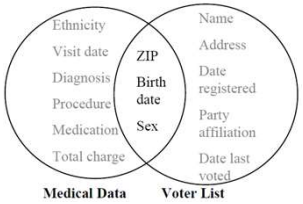
Ανώνυμα δεδομένα: το συχνό λάθος

- 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»

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

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 (or her; but we'll use the male pronoun without loss of generality) 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"*

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Η εποχή των «μεγάλων δεδομένων» (Big Data)

Volume

- 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)

Velocity

- Όχι πλέον μόνο στατικά αλλά και δυναμικά δεδομένα (εκατομμύρια «κλικ» των χρηστών ανά δευτερόλεπτο, τα οποία «αντανακλούν» π.χ., τις συνήθειές τους)

Variety

- Αριθμοί, εικόνες, ήχος, βίντεο, 3D δεδομένα, γεωχωρικά δεδομένα, δεδομένα από κοινωνικά δίκτυα,....

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

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

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

- 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
- Άρα, αν λάβουμε υπόψη όλα τα μέσα τα οποία είναι ευλόγως πιθανό ότι θα χρησιμοποιηθούν για την άμεση ή έμμεση εξακρίβωση της ταυτότητας του φυσικού προσώπου, χρήζει πολύ μεγάλης προσοχής ο χαρακτηρισμός δεδομένων ως ανώνυμων

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Θεσμικό πλαίσιο – Ο νέος Κανονισμός (ΕΕ) 2016/679 (GDPR)

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Προσωπικά Δεδομένα
Το σημερινό θεσμικό πλαίσιο

- Ο ν. 2472/1997 ενσωματώνει την Οδηγία 95/46/ΕΚ
- Προσωπικά δεδομένα: κάθε πληροφορία που αναφέρεται σε φυσικό πρόσωπο («υποκείμενο των δεδομένων»), του οποίου η ταυτότητα είναι γνωστή ή μπορεί να προσδιοριστεί άμεσα ή έμμεσα, ιδίως βάσει αριθμού ταυτότητας ή βάσει ενός ή περισσότερων συγκεκριμένων στοιχείων που το χαρακτηρίζουν από φυσική, βιολογική, ψυχολογική, οικονομική, πολιτιστική ή κοινωνική άποψη
- Με απλά λόγια: κάθε πληροφορία που αναφέρεται σε φυσικό πρόσωπο του οποίου η ταυτότητα είναι γνωστή ή δύναται να εξακριβωθεί
  - το όνομά μας
  - η διεύθυνση (ταχυδρομική αλλά και ηλεκτρονική – email) και το τηλέφωνό μας,
  - τα ενδιαφέροντα και οι απόψεις μας
  - η εικόνα μας (φωτογραφία/video)
  - η διεύθυνση διαδικτύου (IP διεύθυνση) από την οποία «σερφάρουμε»
  - οι σελίδες που επισκεπτόμαστε και τα δεδομένα που παράγουμε στο browser ή το mobile app
  - η θέση της τερματικής συσκευής μας
  - ...

Προσωπικά Δεδομένα
Το επικείμενο θεσμικό πλαίσιο

- Ο Κανονισμός (ΕΕ) 2016/679 (General Data Protection Regulation - GDPR) θα τεθεί σε εφαρμογή σε όλα τα Κράτη Μέλη στις 25 Μαΐου 2018, αντικαθιστώντας την 95/46/ΕΚ
- Προσωπικά δεδομένα: κάθε πληροφορία που αφορά ταυτοποιημένο ή ταυτοποιήσιμο φυσικό πρόσωπο («υποκείμενο των δεδομένων»)· το ταυτοποιήσιμο φυσικό πρόσωπο είναι εκείνο του οποίου η ταυτότητα μπορεί να εξακριβωθεί, άμεσα ή έμμεσα, ιδίως μέσω αναφοράς σε αναγνωριστικό στοιχείο ταυτότητας, όπως όνομα, σε αριθμό ταυτότητας, σε δεδομένα θέσης, σε επιγραμμικό (online) αναγνωριστικό ταυτότητας ή σε έναν ή περισσότερους παράγοντες που προσιδιάζουν στη σωματική, φυσιολογική, γενετική, ψυχολογική, οικονομική, πολιτιστική ή κοινωνική ταυτότητα του εν λόγω προσώπου
- Ουσιαστικά, δεν υπάρχει διαφορά στον ορισμό
  - Ιδιαίτερη επισήμανση σε ειδικού τύπου αναγνωριστικά

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.

Τι νέο φέρει ο GDPR

- Και στον GDPR αναφέρεται ρητώς ότι το θεσμικό πλαίσιο προστασίας προσωπικών δεδομένων δεν εφαρμόζεται σε ανώνυμα δεδομένα
- Και στον GDPR αναφέρεται ρητώς ότι ο χαρακτηρισμός δεδομένων ως ανώνυμων «χρήζει προσοχής»
  - Επεξηγείται ωστόσο περισσότερο το πώς αξιολογούνται τα μέσα που μπορούν να χρησιμοποιηθούν για την αναγνώριση ενός προσώπου ως «ευλόγως πιθανό να χρησιμοποιηθούν»
- Όμως:
  - 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.*

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Η σημασία της ψευδωνυμοποίησης

- Η ίδια ερμηνεία ως προς τα ψευδωνυμοποιημένα δεδομένα ίσχυε και βάσει των διατάξεων της Οδηγίας 95/46/ΕΚ, απλά πλέον διατυπώνεται με σαφήνεια
  - Άρα, η ψευδωνυμοποίηση δεν συνιστά ανωνυμοποίηση
    - Άλλωστε, σε διάφορα σημεία ο GDPR αναφέρει, ως κίνδυνο, την «άρση της ψευδωνυμοποίησης»
- Εν τούτοις, ο GDPR προκρίνει τη χρήση της ψευδωνυμοποίησης για περαιτέρω διασφάλιση της ασφάλειας της επεξεργασίας και της προστασίας των θεμελιωδών δικαιωμάτων
- Η λέξη «ψευδωνυμοποίηση» εμφανίζεται περί τις 15 φορές εντός του GDPR, σε διάφορες εκφάνσεις
  - Πιθανή «κατάλληλη εγγύηση» για:
    - *Επεξεργασία δεδομένων για άλλο σκοπό από αυτόν για τον οποίο έχουν αρχικώς συλλεγεί, η οποία δεν βασίζεται στη συγκατάθεση του προσώπου, προκειμένου να εξακριβωθεί κατά πόσον η επεξεργασία αυτή είναι συμβατή με τον αρχικό σκοπό (άρ. 6, παρ. 3)*
    - *Διασφάλιση της προστασίας των δεδομένων ήδη από το σχεδιασμό (άρ. 25, παρ. 1)*
    - *Ασφάλεια της επεξεργασίας (άρ. 30, παρ. 1)*
    - *Επεξεργασία για σκοπούς αρχειοθέτησης προς το δημόσιο συμφέρον ή σκοπούς επιστημονικής ή ιστορικής έρευνας ή στατιστικούς σκοπούς (άρ. 89, παρ. 1)*
  - Ενθάρρυνση για να λαμβάνεται υπόψη στους κώδικες δεοντολογίας *(άρ. 40, παρ. 2)*

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Ειδικότερες περιπτώσεις που άπτονται ανωνυμοποίησης/ψευδωνυμοποίησης

- Ο βαθμός δυσκολίας της δυνατότητας άρσης της ψευδωνυμοποίησης μπορεί να επηρεάσει στον προσδιορισμό του εάν ένα περιστατικό παραβίασης δεδομένων πρέπει να ανακοινωθεί στα υποκείμενα αυτών
  - Άρθρο 34: Μία τέτοια ανακοίνωση γίνεται όταν «η παραβίαση δεδομένων προσωπικού χαρακτήρα ενδέχεται να θέσει σε υψηλό κίνδυνο τα δικαιώματα και τις ελευθερίες των φυσικών προσώπων»
- Άρθρο 16: Όταν ο υπεύθυνος επεξεργασίας μπορεί να αποδείξει ότι δεν είναι σε θέση να εξακριβώσει την ταυτότητα του υποκειμένου των δεδομένων, ο υπεύθυνος επεξεργασίας ενημερώνει σχετικά το υποκείμενο των δεδομένων.
  - Στις περιπτώσεις αυτές, τα άρθρα 15 ως 20 δεν εφαρμόζονται, εκτός εάν το υποκείμενο των δεδομένων (…) παρέχει συμπληρωματικές πληροφορίες που επιτρέπουν την εξακρίβωση της ταυτότητάς του.
    - Δικαιώματα πρόσβασης / διόρθωσης / διαγραφής / περιορισμού επεξεργασίας/ φορητότητας δεδομένων

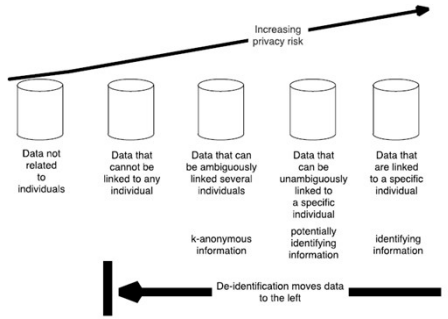
Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Τελικά;

- Με τον GDPR, καθίσταται πιο «εμφατική» η ανάγκη εξέτασης της ψευδωνυμοποίησης ως μέσο το οποίο μπορεί να περιορίσει τους κινδύνους της επεξεργασίας
- Η ψευδωνυμοποίηση ωστόσο σαφώς δεν μπορεί να συνιστά – όσο και αν περιορίζονται οι κίνδυνοι - ανωνυμοποίηση
  - Οι δυσκολίες επίτευξης πλήρους ανωνυμοποίησης παραμένουν
  - Κάθε τεχνική ανωνυμοποίησης ουσιαστικά συντείνει περισσότερο στον περιορισμό των κινδύνων και όχι στο χαρακτηρισμό των δεδομένων ως ανώνυμων
- Για κάθε τεχνική ανωνυμοποίησης/ψευδωνυμοποίησης, θα πρέπει να εξετάζεται η αποτελεσματικότητά της ως προς τους κινδύνους που παραμένουν, αναφορικά με την προστασία προσωπικών δεδομένων
  - Η εξέταση αυτή μπορεί να γίνεται στο πλαίσιο μίας μελέτης αντικτύπου ως προς την προστασία προσωπικών δεδομένων
  - Κρίσιμος παράγοντας είναι η ορθή αξιολόγηση των κινδύνων που παραμένουν

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

«Στάδια» ανωνυμοποίησης



S. L. Garfinkel, "**De-Identification of Personal Information"**, NIST Internal Report 8053, 2015

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

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

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Ανωνυμοποίηση στην πράξη - Διαχείριση αναγνωριστικών

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Η «ευρεία» έννοια του αναγνωριστικού

- Ως προσωπικά δεδομένα θεωρούνται πολύ περισσότερα δεδομένα από ό,τι ίσως φανταζόμαστε
- Παράδειγμα: Τα αναγνωριστικά μιας «έξυπνης» κινητής συσκευής Android

Identifier	Description	Attribute
GAID	User-resettable 32-digit alphanumeric identifier	Pseudonymous
Android ID	64-bit number randomly generated when device is set up for the first time [5]	Semi-permanent
IMEI	15-digit decimal identifier representing GSM or LTE device	Permanent
IMSI	15-digit decimal identifier representing mobile subscriber identity	Permanent
MAC address	48-bit number assigned to the device's Wi-Fi network interface	Permanent

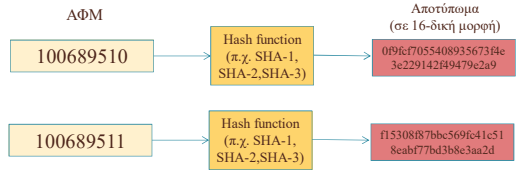
Πηγή: S. Son, D. Kim and V. Shmatikov, "What Mobile Ads Know About Mobile Users", NDSS 2016

- Δύνανται - σε συνδυασμό με άλλες πληροφορίες - να οδηγήσουν σε πλήρη ταυτοποίηση
- Πολλές «έξυπνες» εφαρμογές συλλέγουν το Google Advertising ID (GAID)
  - Συμπεριλαμβανομένων των λεγόμενων ανώνυμων κοινωνικών δικτύων
  - Ακόμα κι αν δεν συλλέγουν στοιχείο ταυτοποίησης του χρήστη, το GAID μπορεί – έστω και υπό προϋποθέσεις – να οδηγήσει σε ταυτοποίηση του χρήστη
    - Π.χ. μπορεί να συλλεγεί και από εφαρμογές που συλλέγουν επίσης και κάποιο αναγνωριστικό του χρήστη (π.χ. e-mail)
  - Ένα είδους ψευδωνυμοποίησης και όχι ανωνυμοποίησης

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

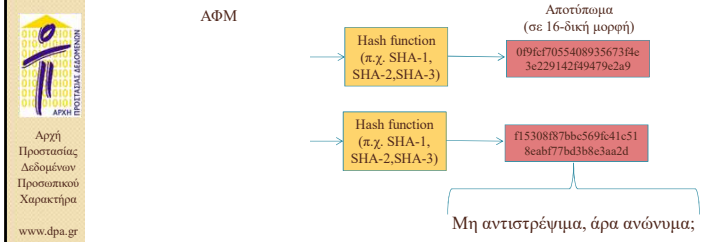
Τι δεν είναι ανωνυμοποίηση

- Κάθε μετασχηματισμός ενός αναγνωριστικού σε μία νέα, ακατάληπτη μορφή, δεν πρέπει να θεωρείται ανωνυμοποίηση αν υπάρχει περίπτωση, από την ακατάληπτη αυτή μορφή, να συμπεράνει/υπολογίσει κανείς την τιμή του αναγνωριστικού.
- Παράδειγμα: Πολλοί υπεύθυνοι επεξεργασίας συλλέγουν το «αποτύπωμα» (**hash value**) ενός μοναδικού αναγνωριστικού, θεωρώντας ότι δεν καθίσταται εφικτή η εύρεση του αρχικού αναγνωριστικού (και, άρα, αναγνώριση του προσώπου) αφού μία κρυπτογραφική συνάρτηση κατακερματισμού (**hash function**) είναι μη αναστρέψιμη.
  - Έτσι επιτυγχάνουν την ιδιότητα «ίδιο άτομο – ίδιο (ακατάληπτο) αναγνωριστικό»



Τι δεν είναι ανωνυμοποίηση

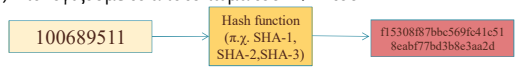
- Κάθε μετασχηματισμός ενός αναγνωριστικού σε μία νέα, ακατάληπτη μορφή, δεν πρέπει να θεωρείται ανωνυμοποίηση αν υπάρχει περίπτωση, από την ακατάληπτη αυτή μορφή, να συμπεράνει/υπολογίσει κανείς την τιμή του αναγνωριστικού.
- Παράδειγμα: Πολλοί υπεύθυνοι επεξεργασίας συλλέγουν το «αποτύπωμα» (**hash value**) ενός μοναδικού αναγνωριστικού, θεωρώντας ότι δεν καθίσταται εφικτή η εύρεση του αρχικού αναγνωριστικού (και, άρα, αναγνώριση του προσώπου) αφού μία κρυπτογραφική συνάρτηση κατακερματισμού (**hash function**) είναι μη αναστρέψιμη.
  - Έτσι επιτυγχάνουν την ιδιότητα «ίδιο άτομο – ίδιο (ακατάληπτο) αναγνωριστικό»



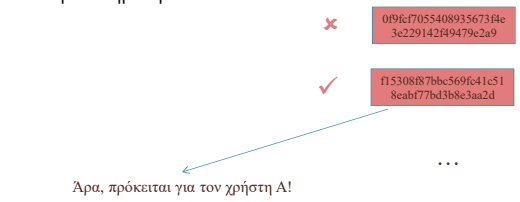
Τι δεν είναι ανωνυμοποίηση (*συνέχεια*)

- Μπορούμε να διαπιστώσουμε αν ο Α με ΑΦΜ 100689511 βρίσκεται στη λίστα;

1. 1) Υπολογίζουμε το αποτύπωμα του ΑΦΜ του Α



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



Βέλτιστη προσέγγιση (ως προς την ανωνυμοποίηση)

- Χρήση κρυπτογραφικής συνάρτησης με μυστικό κλειδί
  - Π.χ. υπολογισμός ενός **Message Authentication Code (MAC)**

- Ακολούθως, το μυστικό κλειδί πρέπει να καταστρέφεται
- Είναι «ισοδύναμο» με το να γίνεται μία τυχαία, μη αντιστρεπτή, αντιστοίχιση του μοναδικού αναγνωριστικού (ΑΦΜ) με μία συμβολοσειρά με χαρακτηριστικά τυχαιότητας

Ανωνυμοποίηση στην πράξη - Διαχείριση ψευδο-αναγνωριστικών

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

- Με τον όρο «δημοσίευση» εννοούμε οποιαδήποτε διαβίβαση/κοινοποίησή του σε τρίτο (π.χ. σε ιατρικό/ερευνητικό κέντρο)
- 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 το ιατρικό κέντρο 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

«Γενίκευση» γνωρισμάτων

- Ως τεχνική ανωνυμοποίησης, μεταβάλλουμε κατάλληλα τις τιμές των πεδίων που είναι ψευδο-αναγνωριστικά, 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

(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» των τιμών των ψευδοαναγνωριστικών

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

Περιγραφή των τιμών γενίκευσης

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 δύο εγγραφών των πινάκων, συγκρίνοντας τα ψευδο-αναγνωριστικά, δεν μπορεί να γίνει
- Ανωνυμία k τάξης (k-anonymity) - Samarati-Sweeney, 1998:
  Ικανοποιείται όταν, σε έναν ανώνυμο πίνακα, το σύνολο των εγγραφών (καταχωρήσεων) με τις ίδιες τιμές στα ψευδο-αναγνωριστικά είναι τουλάχιστον k
- Προφανώς, όσο μεγαλύτερο το k, τόσο ενισχύεται η ανωνυμία
  - Ο ανωτέρω πίνακας είναι ανώνυμος 3ης τάξης

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	Salary
1	13053	28	Russian	20K
2	13068	29	American	20K
3	13068	21	Japanese	40K
4	13053	23	American	40K
5	14853	50	Indian	60K
6	14853	55	Russian	20K
7	14850	47	American	40K
8	14850	49	American	40K
9	13053	31	American	60K
10	13053	37	Indian	60K
11	13068	36	Japanese	60K
12	13068	35	American	60K

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

Ανωνυμία 4ης τάξης

	Non-Sensitive Data			"Sensitive" Data
#	ZIP	Age	Nationality	Salary
1	130**	< 30	*	20K
2	130**	< 30	*	20K
3	130**	< 30	*	40K
4	130**	< 30	*	40K
5	*1485**	*> = 40*	*	*60K*
6	*1485**	*> = 40*	*	*20K*
7	*1485**	*> = 40*	*	*40K*
8	*1485**	*> = 40*	*	*40K*
9	130**	3*	*	60K
10	130**	3*	*	60K
11	130**	3*	*	60K
12	130**	3*	*	60K

Η Akira ανήκει εδώ

Ο Bob ανήκει εδώ

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Ανωνυμία 4ης τάξης

	Non-Sensitive Data			"Sensitive" Data
#	ZIP	Age	Nationality	Salary
1	130**	< 30	*	20K
2	130**	< 30	*	20K
3	130**	< 30	*	40K
4	130**	< 30	*	40K
5	*1485**	*> = 40*	*	*60K*
6	*1485**	*> = 40*	*	*20K*
7	*1485**	*> = 40*	*	*40K*
8	*1485**	*> = 40*	*	*40K*
9	130**	3*	*	60K
10	130**	3*	*	60K
11	130**	3*	*	60K
12	130**	3*	*	60K

Η Akira ανήκει εδώ

Ο Bob ανήκει εδώ

Ο Bob έχει μισθό 60K
- Εξαγωγή συμπεράσματος, παρά την «ανωνυμοποίηση»

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

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)
- 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

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Distinct 3-diversity

	Non-Sensitive Data			«Sensitive» Data
#	ZIP	Age	Nationality	Salary
1	1305*	<= 40	*	20K
2	1305*	<= 40	*	40K
3	1305*	<= 40	*	60K
4	1305*	<= 40	*	60K
5	*1485**	*>= 40*	*	*60K*
6	*1485**	*>= 40*	*	*20K*
7	*1485**	*>= 40*	*	*40K*
8	*1485**	*>= 40*	*	*40K*
9	1306*	<= 40	*	20K
10	1306*	<= 40	*	40K
11	1306*	<= 40	*	60K
12	1306*	<= 40	*	60K

Οι Bob, Akira ανήκουν εδώ

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Τεχνικές ανωνυμοποίησης - Συμπέρασμα

- Η ανωνυμοποίηση είναι εξαιρετικά δύσκολη ως διαδικασία
- Ούτε η τεχνική **L-diversity** μπορεί να τη διασφαλίσει πάντα
  - Π.χ. αν για τον **Bob** είχαμε **L** διαφορετικές μεν τιμές αλλά πολύ «κοντινές» μεταξύ τους, πάλι θα υπήρχε ο κίνδυνος εξαγωγής συμπεράσματος
- Κάθε περίπτωση ανωνυμοποίησης πρέπει να κρίνεται ξεχωριστά, λαμβάνοντας υπόψη όλους τους πιθανούς κινδύνους
  - Η κατανομή των «ευαίσθητων» τιμών, αλλά και οι τιμές αυτές καθ' αυτές, επηρεάζουν δραστικά την αποτελεσματικότητα της ανωνυμοποίησης
- Το πιθανότερο είναι ότι, με αυτές τις τεχνικές, περιορίζουμε μεν τους κινδύνους ως προς την προστασία προσωπικών δεδομένων αλλά δεν καθιστούμε, τελικά, τα δεδομένα ανώνυμα

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

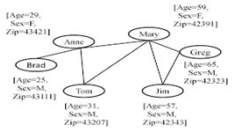
Νέες τεχνολογίες - Γράφοι κοινωνικής δικτύωσης

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Αναπαράσταση κοινωνικών δικτύων με γράφους

Για επιστημονική αξιοποίηση (π.χ. στατιστική ανάλυση) της πληροφορίας που απορρέει από τα κοινωνικά δίκτυα, χρησιμοποιούνται οι λεγόμενοι γράφοι κοινωνικής δικτύωσης (social network graphs)

- Κόμβοι για τη μοντελοποίηση των χρηστών (με αναγνωριστικά και ψευδοαναγνωριστικά)
- Ακμές για τη μοντελοποίηση συνδέσεων
  - Ενδεχομένως με ετικέτες («οικογένεια», «συμμαθητές», «συνάδελφοι» κτλ.)

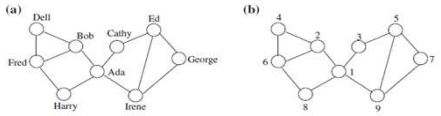


Η απαλοιφή των αναγνωριστικών ή/και η γενίκευση των ψευδοαναγνωριστικών δεν είναι αρκετή για να χαρακτηριστεί ένας τέτοιος γράφος ως ανώνυμος

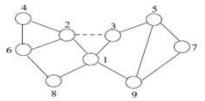
- Ήδη από το 2009, σε έναν ανώνυμο γράφο του Twitter αναγνωρίστηκε περίπου το 1/3 των χρηστών (λογαριασμοί Twitter), συνδυάζοντας πληροφορίες από το Flickr

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Ανωνυμία k βαθμού



- Αν ξέρουμε ότι η Ada έχει 4 φίλους, αναγνωρίζεται αμέσως στον «ανωνυμοποιημένο» γράφο γιατί μόνο ένας κόμβος έχει βαθμό 4.
- Ανωνυμοποίηση γράφου: Προσθέτουμε ή αφαιρούμε ακμές για να επιτύχουμε ανωνυμία k βαθμού (k-degree anonymity) – δηλαδή, για κάθε κόμβο, να υπάρχουν τουλάχιστον άλλοι k-1 με τον ίδιο βαθμό με αυτόν
  - Στόχος: η ελάχιστη δυνατή τροποποίηση του γράφου για την επίτευξη της επιθυμητής ανωνυμίας

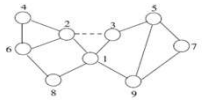


Ανωνυμία βαθμού 2

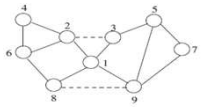
Πηγή: N. Lin and S. J. Das, Applications of k-Anonymity and ℓ-Diversity in Publishing Online Social Networks, Security and Privacy in Social Networks, 2013

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα
www.dpa.gr

Ανωνυμία k γειτνίασης



- Οι κόμβοι 1 και 2 είναι οι μόνοι με βαθμό 4, αλλά έχουν διαφορετική «γειτονιά»
  - Οι τέσσερις φίλοι του κόμβου 2 είναι δύο ζευγάρια φίλων (4-6 και 1-3), ενώ αυτό δεν ισχύει για τους τέσσερις φίλους του κόμβου 1
- Ορίζεται λοιπόν αντίστοιχα η ανωνυμία k γειτνίασης (k-neighborhood anonymity)



Ανωνυμία 2 γειτνίασης

- Το πρόβλημα γίνεται ακόμα πιο σύνθετο όταν οι ακμές έχουν ετικέτες

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Νέες τεχνολογίες - Τεχνολογίες Blockchain

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

"Distributed ledger" και "Blockchain"

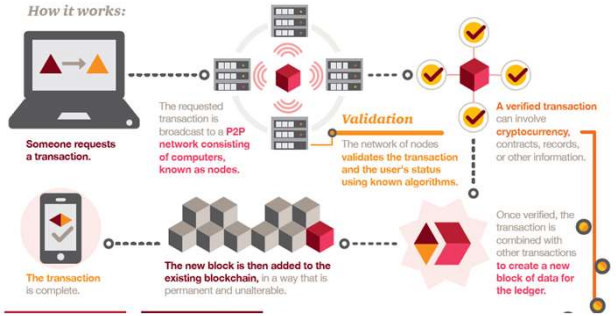
- Ο όρος «distributed ledger» χρησιμοποιείται για να περιγράψει μία οποιουδήποτε τύπου βάση δεδομένων η οποία είναι κατανεμημένη σε διάφορα γεωγραφικά μέρη, όπου όλοι οι χρήστες μπορούν να έχουν πρόσβαση στο σύνολό της και να επιβεβαιώνουν την εγκυρότητα των στοιχείων της
  - Μπορεί να είναι είτε ελεύθερα προσβάσιμη από όλους (permissionless ledger) είτε σε συγκεκριμένους μόνο χρήστες (permissioned ledger)
- Εάν μία τέτοια βάση δεδομένων αποτελείται από ένα σύνολο blocks συγκεκριμένης δομής, συνδεδεμένων το ένα με το άλλο με μία χρονική ακολουθία, τότε αναφερόμαστε σε «αλυσίδα blocks» (blockchain)
  - Άρα, το blockchain αποτελεί υπο-περίπτωση ενός distributed ledger
- Η τεχνολογία blockchain επιτρέπει την πραγματοποίηση «συναλλαγών» χωρίς την ανάγκη ύπαρξης κάποιας έμπιστης τρίτης οντότητας – οι ίδιοι οι χρήστες, βάσει κανόνων, επικυρώνουν τις συναλλαγές
  - Η blockchain τεχνολογία ξεκίνησε με το ψηφιακό νόμισμα Bitcoin, αλλά πλέον έχει διευρυνθεί («ψηφιακά συμβόλαια»)
  - Αναμένεται να έχει δραστικότατο ρόλο στα προσεχή χρόνια

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

(Απλοποιημένη) περιγραφή

"Every informed person needs to know about Bitcoin because it might be one of the world's most important developments." Leon Luow, Nobel Peace prize nominee



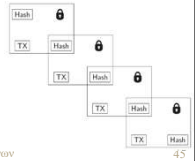
- Οι χρήστες έχουν «ψευδώνυμα»
- Κανένας κόμβος της αλυσίδας δεν μπορεί να διαγραφεί
- Όλοι οι κόμβοι είναι διαθέσιμοι σε όλους τους χρήστες

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Προκλήσεις εν όψει GDPR

- Αν και υπάρχει διάχυτη η αίσθηση ότι η τεχνολογία blockchain παρέχει ανωνυμία, τα δεδομένα ουσιαστικά είναι ψευδωνυμοποιημένα
  - Κάθε χρήστης είναι συνδεδεμένος με ένα μοναδικό κρυπτογραφικό αναγνωριστικό
- Άρα, ο GDPR θα έχει εφαρμογή
- Η έννοια των «μεγάλων δεδομένων» αποκτά ακόμα ευρύτερη ερμηνεία
- Πολλά ανοιχτά ερωτήματα
  - Υπεύθυνος επεξεργασίας;
  - Δικαίωμα στη λήθη;
    - Κανένας κόμβος του blockchain δεν μπορεί να διαγραφεί, με την υπάρχουσα δομή σχεδίασής του
    - Χρειάζονται τροποποίηση δομικά συστατικά της σχεδίασής του;
  - και πολλά ακόμη....



Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Συμπεράσματα

- Η ανωνυμοποίηση είναι εξαιρετικά δύσκολο να επιτευχθεί
  - Δεν υπάρχει απόλυτα εγγυημένη τεχνική ανωνυμοποίησης που να μπορεί να εφαρμοστεί οποτεδήποτε και οπουδήποτε
- Πολλοί θεωρούν εσφαλμένα την ψευδωνυμοποίηση ως ανωνυμοποίηση
  - Η ψευδωνυμοποίηση είναι ένα μέτρο προστασίας δεδομένων – ενδεχομένως αποτελεσματικό σε ορισμένες περιπτώσεις εν όψει του επιδιωκόμενου σκοπού επεξεργασίας – αλλά δεν είναι ανωνυμοποίηση
  - Ειδική αναφορά και στον GDPR
- Τεχνικές ανωνυμοποίησης/ψευδωνυμοποίησης πρέπει να εφαρμόζονται πριν από την ανάλυση των δεδομένων
  - Αναμένεται να προκύψουν ως αποτέλεσμα μιας ορθά εκπονηθείσας μελέτης αντικτύπου ως προς την προστασία προσωπικών δεδομένων
- Η προστασία των προσωπικών δεδομένων δεν θα πρέπει να εκλαμβάνεται ως «εμπόδιο» για την επιστημονική ανάλυση και αξιοποίηση των «μεγάλων δεδομένων», αλλά ως αναπόσπαστο στοιχείο αυτής, από το οποίο ωφελημένοι τελικά είναι όχι μόνο τα φυσικά πρόσωπα αλλά και όσοι πραγματοποιούν την ανάλυση

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr

Ερωτήσεις?

Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα

www.dpa.gr