1η ημέρα Διαλόγου με την Ερευνητική Κοινότητα
Τετάρτη 1 Οκτωβρίου 2025

«Ανιχνευτές Ιστού (Web Crawlers) υποστηριζόμενοι από συστήματα Τεχνητής Νοημοσύνης (το παράδειγμα της αναγνώρισης προσώπου) και η προστασία δεδομένων»

Ανδρέας Κανακάκης, Υπ. Διδάκτωρ, Ερευνητής, Vrije Universiteit Brussel/Université du Luxembourg

VRIJE UNIVERSITEIT BRUSSEL

FUNDAMENTAL RIGHTS RESEARCH CENTRE

UNIVERSITÉ DU LUXEMBOURG

Επισκόπηση Παρουσίασης

ΘΕΜΑΤΙΚΟΙ ΑΞΟΝΕΣ

- **Αυτοματοποιημένη συλλογή δημοσίως διαθέσιμων διαδικτυακών πληροφοριών**
- **Ανιχνευτές Ιστού υποστηριζόμενοι από ΤΝ – το παράδειγμα της αναγνώρισης προσώπου**

Εισαγωγικά ερωτήματα:

- Σε τι χρησιμεύουν αυτά τα εργαλεία και γιατί είναι σημαντικά για τις αρχές επιβολής του νόμου;
- Πώς λειτουργούν στην πράξη;
- Πώς άπτονται ζητημάτων ιδιωτικότητας, προστασίας δεδομένων και νομοθετικής ρύθμισης;

‘Δημόσια Διαθέσιμες Διαδικτυακές Πληροφορίες’

ΟΡΙΣΜΟΣ & ΠΑΡΑΔΕΙΓΜΑΤΑ

Συστατικά Στοιχεία

- Ανεμπόδιστη (τεχνική) προσβασιμότητα χωρίς login, πληρωμή ή ιδιότητα μέλους
- Εντοπίζονται σε surface, deep και dark web
- *Αντικειμενικά* και *υποκειμενικά* αντιληπτές ως «δημόσια διαθέσιμες»

Συναφείς όροι (< υπηρεσίες πληροφοριών):

- OSINF: ακατέργαστα δεδομένα (σχόλια, εικόνες, ψηφιακά ίχνη κ.ά.)
- OSINT: δομημένες, αξιοποιήσιμες πληροφορίες

Πρακτικά Παραδείγματα

- **London Riots (2011)**: αναρτήσεις στα μέσα κοινωνικής δικτύωσης βοήθησαν στον εντοπισμό δραστών
- **Silk Road (2013)**: ανάρτηση σε δημόσιο φόρουμ οδήγησε στη σύλληψη του Ross Ulbricht
- **Επίθεση στο Αμερικανικό Καπιτώλιο (2021)**: πάνω από 90% των συλλήψεων βασίστηκαν σε πληροφορίες από ανοιχτές πηγές

Μέθοδοι Συλλογής Δημόσιων Διαδικτυακών Πληροφοριών

Παραδοσιακές Μέθοδοι

- Παρατήρηση σε πραγματικό χρόνο: παρακολούθηση δραστηριότητας καθώς εξελίσσεται (π.χ. livestreams, σχόλια)
- Χειροκίνητη συλλογή δεδομένων: αναζήτηση με λέξεις κλειδιά, hashtags ή περιήγηση σε μέσα κοινωνικής δικτύωσης
- Γιατί είναι ανεπαρκείς σήμερα:
  - Όγκος, ταχύτητα και ποικιλία δεδομένων υπερβαίνουν τις ανθρώπινες δυνατότητες
  - Δεν συμβαδίζουν με τα σύγχρονα ψηφιακά περιβάλλοντα.

Αυτοματοποιημένες Μέθοδοι

Web Crawling	Web Scraping
• Συστηματική περιήγηση στο διαδίκτυο για ανακάλυψη περιεχομένου • Ακολουθεί συνδέσμους από “seed” URLs, χαρτογραφεί δομές • Παράδειγμα: Googlebot, GPTBot • Εστίαση: ανακάλυψη, όχι εξαγωγή	• Εξαγωγή συγκεκριμένων δεδομένων (π.χ. posts, metadata, εικόνες) • Δόμηση σε αξιοποιήσιμη μορφή (π.χ. βάσεις δεδομένων) • Εστίαση: ανάκτηση και επεξεργασία

«Έξυπνες» Μέθοδοι Συλλογής Δεδομένων

Από αυτοματοποιημένη σε «έξυπνη» συλλογή

- Εργαλεία ΤΝ τροποποιούν την επεξεργασία, ερμηνεία και χρήση συλλεγέντων δεδομένων
- Βασίζονται είτε στην ανίχνευση (δεδομένα εκπαίδευσης) είτε στην «απόξυση» (στατικά αποθετήρια)
- Ειδοποιός διαφορά: στάδιο επεξεργασίας → από συσσώρευση δεδομένων στην παραγωγή γνώσης

Μεγάλα Γλωσσικά Μοντέλα (LLMs)

- Εκπαιδευμένα σε «ανιχνευθέν» υλικό
- Δημιουργούν απαντήσεις με βάση πιθανολογική κατανόηση προτροπών χρηστών
- Δυνατότητες: σύνοψη περιεχομένου, αποκωδικοποίηση γλώσσας φόρουμ (π.χ. αργκό μαύρων αγορών), σκιαγράφηση προφίλ υπόπτων

Τεχνολογία Αναγνώρισης Προσώπου

- αμφ. εάν πρόκειται για crawling ή scraping
- Παραδείγματα εμπορικών εργαλείων: PimEyes vs. ClearviewAI
- Crawling σε πραγματικό χρόνο ή στατική βάση δεδομένων΄

Παράγοντες Επέμβασης: Είδος Πληροφοριών

Δημόσιες = απροστάτευτες?

- Δημόσια διαθεσιμότητα ≠ άρση προστασίας βάσει άρθρου 8 ΕΣΔΑ και άρθρων 7–8 ΧΘΔΕΕ
- Ανάρτηση στο διαδίκτυο ≠ παραίτηση από δικαιώματα ιδιωτικότητας και προστασίας δεδομένων → άρση απαγόρευσης άρθρου 10 LED (9 GDPR), βλ. άρθρο 10 περ. γ' LED (9 παρ. 2 περ. ε' GDPR) μεν, αλλά όχι άρση προστασίας.
- Σημασία της φύσης των δεδομένων:
  - Προσωπικά → κάθε πληροφορία που αφορά ταυτοποιημένο ή ταυτοποιήσιμο φυσικό πρόσωπο
  - Ευαίσθητα → π.χ. φυλή, υγεία, πολιτικές απόψεις → σοβαρότερη επέμβαση
  - Ψευδωνυμοποιημένα → παραμένουν προσωπικά, με κίνδυνο επαναταυτοποίησης

- Δημόσια διαθέσιμα αλλά ιδιωτικά και προσωπικά
- Βαθμός επέμβασης < είδος δεδομένων + εγγύτητα στην ιδιωτική ζωή

Παράγοντες Επέμβασης: Τύπος Μεθόδου

Επέμβαση βάσει μεθόδου

Παραδοσιακές Μέθοδοι Συλλογής Δεδομένων

- **Πρόχειρη – μεμονωμένη συλλογή** = μικρή/ ελάχιστη επέμβαση (arg. '*reasonable expectation to privacy*', βλ. ECtHR, *P.G. and J.H. v. UK,* παρ. 57 και *Perry v. UK,* παρ. 37)
- **Συστηματική συλλογή + αποθήκευση** = επέμβαση (ECtHR, *Segerstedt-Wiberg and Others v. Sweden* παρ. 71-72; *Catt v. the United Kingdom*, παρ. 93)
  - **Αποθήκευση *μόνο*** = επέμβαση ανεξάρτητα από επακόλουθη χρήση (E*CtHR, Amann v. Switzerland*, [αρ. 69; *Gardel v. France*, παρ. 58–62)
  - **Συστηματική συλλογή *μόνο*** = ? (arg. κλίμακα και ταχύτητα εντείνουν την επίδραση + κανόνες προστασίας δεδομένων ισχύουν ανεξάρτητα από τη μετέπειτα χρήση).

Παράγοντες Επέμβασης: Τύπος Μεθόδου

Επέμβαση βάσει μεθόδου

Αυτοματοποιημένη Συλλογή Δεδομένων

Τεστ 4 Παραγόντων (ECtHR, *S. and Marper v. UK*, παρ. 67):

1. **Πλαίσιο συλλογής: δυνατότητα συνεχούς κι αδιάκριτης λειτουργίας** (καθεστώς κεκαλυμμένης μαζικής παρακολούθησης+ chilling effects), πρβλ. ECtHR, *Big Brother Watch and Others v. UK*, παρ. 225; CJEU, *Digital Rights Ireland*, παρ. 37.
2. **Φύση αρχείων:** διαβαθμισμένη προσέγγιση ανάλογα με το είδος των δεδομένων (λ.χ. εικόνες προσώπων ≠ DNA, αλλά η δημόσια διαθεσιμότητα μειώνει – όχι αναιρεί – τον αντίκτυπο)
3. **Χρήση/επεξεργασία:** προληπτική κι αδιάκριτη αποθήκευση + αυτοματοποίηση επεξεργασίας → σοβαρή επέμβαση
4. **Κτηθέντα αποτελέσματα:** πέρα από ουδέτερη ταυτοποίηση → δημιουργία «μωσαϊκού ιδιωτικής ζωής» → ενισχυμένη επέμβαση

➢ Ανάγκη για αυξημένες εγγυήσεις προστασίας (*S. and Marper v. UK*, παρ. 103; *M.K. v. France*, παρ. 35)

Καταληκτικές Σκέψεις

ΕΙΔΟΣ ΝΟΜΟΘΕΤΙΚΗΣ ΡΥΘΜΙΣΗΣ

Αποδομώντας τον «μύθο»

- Δημόσια διαθέσιμα δεδομένα ≠ εξ υπαρχής ελάχιστη επέμβαση
- Έμφαση στη μέθοδο (ιδιαίτερα < χρήση αυτοματοποιημένων «έξυπνων» εργαλείων)

Κανονιστικές προεκτάσεις

- Μέθοδοι μικρής επέμβασης (παραδοσιακές): γενική κανονιστική βάση + συμμόρφωση με κανόνες προστασίας δεδομένων
- Αυτοματοποιημένα εργαλεία (crawlers, scrapers, έξυπνα συστήματα): απαιτείται ειδική κανονιστική βάση, ορίζουσα π.χ. τα χρονικά όρια ή τις συνθήκες εφαρμογής (ένταλμα & ανεξάρτητη εποπτεία)

Τρία επίπεδα ρύθμισης:

- Άμεση εφαρμογή δικαίου προστασίας δεδομένων
- Μικρές επεμβάσεις → γενική κανονιστική βάση + κανόνες προστασίας δεδομένων
- Συστηματική/ αυτοματοποιημένη/ «έξυπνη» συλλογή δεδομένων → ειδική κανονιστική βάση + δικονομικές εγγυήσεις

Ερωτήσεις;

Ευχαριστώ θερμά για την προσοχή σας!

Η ολοκληρωθείσα παρουσίαση αποτελεί ερευνητικό αποτύπωμα στο πλαίσιο του προγράμματος PROMODE (μτφ. *Εκσυγχρονίζοντας τους Κώδικες Ποινικής Δικονομίας σχετικά με τις Ψηφιακές Μεθόδους Έρευνας*), υποστηριζόμενου από τα Research Foundation – Flanders (FWOAL1089) και Luxembourg National Research Fund (INTER/FWO/22/17204313).