1° Convegno Nazionale

CYBER RISK IN SANITÀ

MILANO
Auditorium Giovanni Testori
Palazzo Lombardia
12 aprile 2017

Piero Andrea Bonatti
Università di Napoli Federico II

La protezione dei dati nell'era del Web Semantico: Realtà e prospettive

Le tendenze in Europa e USA

- Regolamentazione privacy meno vincolante negli USA che in Europa
  - Trump: verso una deregolamentazione parziale dell'uso delle informazioni raccolte on-line
- In Europa: la GDPR (General Data Protection Regulation) entrerà in vigore nel maggio 2018
  - Combinata alla direttiva ePrivacy
- Simultaneamente la ricerca europea sulla privacy cerca di «compensare» i vincoli introdotti dalla normativa

Strategie diverse, scopi simili

- L’obiettivo in entrambi i casi è la competitività delle aziende
  - Attraverso capacità di innovazione
  - Supportata dall’utilizzo intensivo dei dati disponibili
- In Europa si vuole perseguire senza sacrificare la privacy
- L’obiettivo di innovazione è evidente in diversi documenti (slides successive)

Strategie diverse, scopi simili

- Rapporto ENISA (European Union Agency for Network and Information Security)

  Privacy by design in big data (2015)

- Superare il conflitto "big data *versus* privacy" per arrivare a "big data *with* privacy"
  - Preservando l'*utilità* dei dati ovvero permettendo:
  - Linkability & Composability
  - Analytics

Strategie diverse, scopi simili

- La *challenge* del bando H2020 ICT-18-2016 (Big data PPP: privacy-preserving big data technologies) cita:
  - Businesses are often unsure about how to deal with the data
    - Timore di conseguenze legali in caso di uso improprio
    - Timore di perdita di *reputazione*
    - Potenziali danni economici molto gravi
  - This data is of particularly high value
    - Ma le aziende spesso rinunciano a utilizzarli !
  - Citizens, consumers […] often feel that they have no control over the use of their personal data
  - The resulting lack of confidence undermines efficient and legitimate data sharing and value creation for agreed purposes

Strategie diverse, scopi simili

- La *challenge* del bando H2020 ICT-18-2016 (Big data PPP: privacy-preserving big data technologies) è:
  - to develop technologies […] for empowering
  - the data subjects to understand and be informed of (and […]control) the use of their personal data
  - and the entrepreneurs to develop and run their data driven business

Evitare conseguenze paradossali

- Non solo i vincoli sull'uso dei dati personali possono rendere le aziende europee meno competitive
  - Spingendole a *non* utilizzare i dati disponibili
- Paradossalmente i cittadini europei – non trovando in Europa i servizi che cercano – sarebbero spinti verso aziende extraeuropee
  - Con poche / nulle garanzie sull'uso dei dati
  - Vanificando nei fatti la GDPR
- Si è compreso che occorre un delicato compromesso tra
  - Protezione della privacy
  - Utilizzo proficuo dei dati

L’esempio dei dati medici

- Negli USA soggetti influenti (compresi Microsoft ed Apple) forniscono servizi e app per health & fitness
  - Non soggetti a HIPAA !
- Possono manipolare disinvoltamente
  - Dati biomedici
  - PII (Personally Identifiable Information)
  - Persino interi record medici
- Mentre in Europa non si può fare praticamente nulla con dati simili

Gli approcci tradizionali alla privacy

- Focalizzati sulla *anonimizzazione*
- Principali metodi di anonimizzazione:
  - k-anonimity ed estensioni (l-diversity, t-closeness ecc.)
  - Differential privacy e metodi derivati
- La prima famiglia toglie dettagli ma l'informazione è corretta
- La seconda famiglia perturba i dati per offuscarli

Limiti dell'anonimizzazione

- L'efficacia dell'anonimizzazione è messa a rischio da
  - Abbondanza di sorgenti di informazioni con cui collegare e arricchire i dati
  - Progressi dell'Intelligenza Artificiale (AI)
    - Data mining
    - Automated reasoning
    - Applicabili su larga scala
  - Entrambi permettono di ricostruire molta informazione cancellata o offuscata
- Semantic web ≈ WWW + AI
- Ma anche Big Data + AI

Limiti dell'anonimizzazione

- K-anonymity e sue estensioni sono particolarmente vulnerabili ai collegamenti con ulteriori fonti di informazione
- Le varianti di differential privacy sono più robuste (ma sempre sensibili) rispetto all'uso di altre sorgenti di dati, ma…
- Spesso tolgono valore/utilità ai dati
- E richiedono di restringere a priori le modalità di interrogazione dei dati (*query*)
  - In conflitto con l'aggiunta di nuovi utilizzi
  - In conflitto con analytics per innovazione
  - Molti ambiti applicativi non consentono restrizioni a priori

[cf. Privacy by Design in Big Data. ENISA report 2015]

Oltre l’anonimizzazione

- Dato che
  1. L’anonimizzazione *effettiva* è pressochè impossibile, tende a rendere i dati inutilizzabili ed è difficile da applicare
  2. Le aziende sono interessate a utilizzare legalmente i dati per evitare sanzioni e perdita di reputazione/clienti
     - Al punto che attualmente rinunciano a usarli
  3. Gli utenti rilascerebbero più facilmente i propri dati se potessero mantenerne il controllo
- Conviene esplorare soluzioni **non** basate unicamente sull’anonimizzazione e sull’offuscamento dei dati
- Facendo leva sui punti 2 e 3
- E sul supporto delle leggi

Trasparenza e consenso l'approccio di SPECIAL

- Progetto H2020 SPECIAL
  - Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance
  - Bando *Big Data PPP: privacy-preserving Big Data technologies* (ICT-18-2016)
  - Inizio: Gennaio 2017
- Fa leva su:
  - Necessità dei *data processor* europei di rispettare le norme (GDPR)
    - *Voluntary compliance*
  - Desiderio dei *data subject* di riprendere il controllo dei propri dati senza rinunciare ai servizi del mondo digitale

Trasparenza e consenso l'approccio di SPECIAL

- I pilastri tecnici di SPECIAL
  1. Un registro di eventi di manipolazione dei dati
     - Che traccia raccolta, elaborazioni e trasmissioni dei dati
  2. Rappresentazione *machine understandable* di «politiche»
     - Che rappresentano vincoli di utilizzo
     - ad es. GDPR, user privacy preferences, company policies
     - Per automatizzare i controlli di conformità a tutti i vincoli di utilizzo
  3. *Sticky policies*
     - I vincoli di utilizzo restano associati ai dati, anche quando vengono trasmessi
  4. Dashboard per i soggetti dei dati
     - Per facilitare il tracciamento dei propri dati e il loro uso
     - Per uniformare le comunicazioni tra *data subjects* e *data processors*

Trasparenza e consenso l'approccio di SPECIAL

- Con il registro delle manipolazioni dati e le politiche *machine understandable* i *data processors* potranno
  - Automatizzare le verifiche di conformità delle proprie procedure
  - Dimostrare il corretto utilizzo dei dati a soggetti e controllori
- Mediante la dashboard
  - Il registro sarà reso trasparente e intelligibile
  - I *data subjects* potranno esercitare i diritti di accesso, rettifica e cancellazione
  - I *data processors* potranno chiedere il consenso per il riutilizzo dei dati (o l'elaborazione di nuovi dati)

Quali garanzie oltre alle sanzioni previste?

- Il registro delle manipolazioni dati contiene asserzioni *non ripudiabili né modificabili*
  - I soggetti non possono negare di aver dato il consenso a certe condizioni di utilizzo
  - I *data processor* non possono negare di aver ricevuto i dati sotto certi vincoli di utilizzo
  - I trasferimenti di dati non possono essere negati unilateralmente
- Tali asserzioni costituiscono una evidenza abbastanza forte da essere utilizzata in caso di contenzioso legale a tutela di tutte la parti
  - Garantita da firma digitale e protocolli *fair exchange*

Il ruolo dei *Linked Data* in SPECIAL

- I linked data sono la base del web semantico
  - Rappresentano *conoscenza*
- Data format: triple (*soggetto, predicato, oggetto*)
  - Soggetti e oggetti sono URI
  - Formati XML standard: RDF, OWL
- Ideali per rappresentare vincoli come le «politiche»
  - E loro diversi usi: verifiche di ottemperanza, conformità con le leggi…
  - Sono tutti forme di *ragionamento* sulle politiche basate su una unica rappresentazione
- Facili da integrare con i sistemi esistenti (basta associare URI a risorse e permetterne le dereferenziazione)

La scommessa di SPECIAL

- Più trasparenza → più fiducia e controllo
- Più fiducia e controllo → più consenso
- Più consenso e controllo → più innovazione nel rispetto della privacy

Visioni future

- Immaginiamo che una infrastruttura di trasparenza «alla SPECIAL», sia richiesta per legge
  - E obbligatorio mantenere provenienza (certificata) dei dati
- Potrebbe supportare indagini su abusi dei dati
  - Facilitando il tracciamento dei rilasci illegali
  - Introducendo incentivi al rispetto delle leggi e della volontà dei soggetti dei dati
  - La cui efficacia è quantificabile mediante analisi basate su teoria dei giochi

Vedasi proof of concept in
[Bonatti et al., Data Usage Workshop, Security & Privacy 2013]

Fine presentazione