Privacy e rispetto dei dati sensibili ai tempi dei Big Data

Giulia Bonansinga
LM in Informatica Umanistica, matricola 442957
Relazione finale per il Seminario di Cultura Digitale

1 Introduzione

La tecnologia pervade un numero sempre maggiore di attività umane, rendendo disponibili enormi quantità di dati provenienti dalle fonti più disparate. Moltissime azioni compiute sul Web possono essere registrate e analizzate, diventando parte attiva della conoscenza veicolata su Internet; per esempio, le query con cui si interrogano i motori di ricerca possono essere utilizzate per raffinare la performance degli algoritmi per *Information Retrieval* e il ranking assegnato alle pagine web. Spostando lo sguardo sui social network o sui quotidiani on-line, è facile individuare gli argomenti che stanno catalizzando l'attenzione degli utenti e svolgere un'analisi di *opinion mining* per ottenere una sintesi dell'opinione pubblica sui temi caldi del momento. Ancora, contenuti simili possono essere aggregati e proposti ai navigatori sfruttando i metadati o gli *hashtag* che accompagnano le condivisioni di articoli, immagini e video, magari facendo uso delle informazioni di geolocalizzazione presenti negli status dei principali social network.

L'esplosione di dati digitali è naturalmente correlata all'aumento di utenti e delle loro attività sul web: l'infografica in Fig. 1 offre una panoramica dell'attività sul Web in un minuto; l'immagine riporta dati dell'aprile 2015, ma i numeri sono in costante crescita. Al contempo, diventa sempre più sostenibile non solo creare, ma anche "immagazzinare" grandi quantità di dati; si stima che nel 2020 un petabyte (corrispondente a 10^3 terabyte) costerà solo 4$, come riporta la Forrester Research, un'azienda specializzata nelle ricerche di mercato sull'impatto della tecnologia sulle attività umane [12].

Il paragrafo 2 riassume brevemente l'intervento sulla *Social Data Science* presentato da Pedreschi nell'ambito del Seminario di Cultura digitale [20] e introduce gli argomenti oggetto di questa relazione, ossia il rischio per la privacy che si pone svolgendo attività sul web e quali misure si possano prendere per la protezione dei dati sensibili.

Il paragrafo 3 riassume la normativa europea in tema di privacy, con un occhio di riguardo a come il problema della privacy è percepito e trattato in Italia.

Nel paragrafo 4 si offre una breve introduzione al *data mining*, l'insieme di tecniche e approcci per l'estrazione di conoscenza da grandi quantità di dati. Il paragrafo 5 approfondisce il problema del trattamento e della protezione delle informazioni sensibili nel data mining, sia che queste siano visibili nei dati di partenza, sia che siano deducibili e riconducibili ai singoli individui nella conoscenza da essi estratta. Vengono anche presentate le principali famiglie di approcci attualmente utilizzate per il *Privacy Preserving Data Mining*, insieme a una descrizione degli algoritmi più utilizzati. Infine, nel paragrafo 6 si cerca di trarre delle conclusioni e di guardare con una prospettiva più consapevole al problema della privacy su Internet.

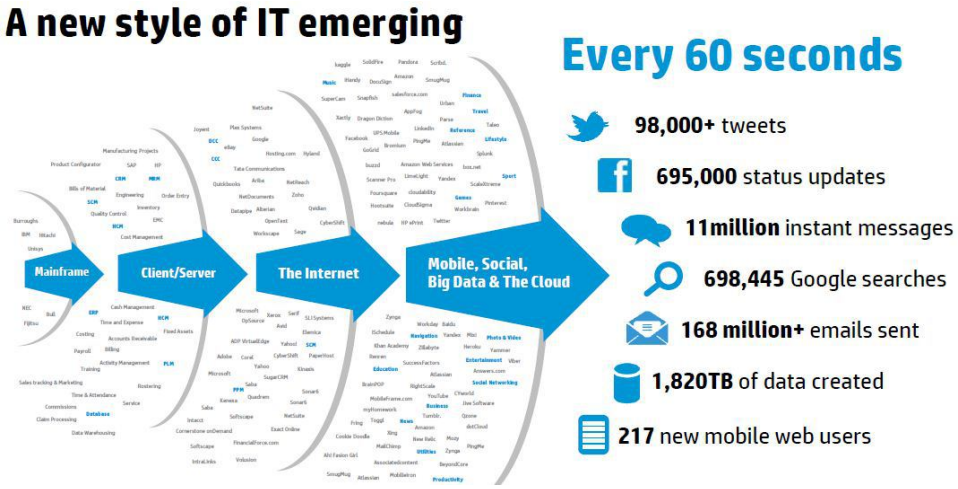

Figura 1: Quanti dati si generano nell'arco di un minuto? [Fonte: `http://31december2099.wordpress.com/`]

2 L'avvento di una *Social Data Science*

Per *Big Data* si intendono dei set di dati, *dataset*, talmente estesi che è impossibile elaborarli ed estrarne conoscenza con le tecniche di analisi tradizionali. La dimensione eccezionale è la caratteristica definitoria di tali dataset, ma con Big Data ci si riferisce anche a dati di grande complessità e variabilità interna, prodotti a ritmi impareggiabili. Accogliendo l'avvento di una *Social Data Science*, Pedreschi parla di una società "misurabile", in cui ogni attività umana lascia tracce digitali e può essere studiata in modo quantitativo [20]. In questo scenario, il vero punto di rottura consiste nel fatto che questi dati sono disponibili anche a un'utenza non specializzata.

Fino allo scorso decennio, per esempio, dati sulle tendenze della popolazione - in termini di fenomeni, mode, opinioni - erano fruibili al grande pubblico solo tramite statistiche generali, che descrivevano solo un numero limitato di fenomeni selezionati ed erano veicolate da enti appositi. Al giorno d'oggi, invece, le API (*Application Program Interface*) dedicate e gli aggregatori di dati permettono a qualsiasi utente di trovare, consultare e utilizzare dati di (quasi) ogni tipo.

Pedreschi riporta numerosi esempi che illustrano le innumerevoli potenzialità offerte dall'analisi di "dati sociali"; per esempio, un festival musicale a Parigi diventa un'opportunità per uno studio della mobilità, sia sotto forma di spostamenti individuali che collettivi. Le applicazioni possibili sono numerose: controllo del traffico, integrazione con eventi contemporanei in aree adiacenti, pianificazione della localizzazione di servizi e punti di ristoro, eccetera.

Lo spazio di investigazione non si limita però a dati e scopi "sociali": il trattamento dei Big Data permette di ricostruire e predire crisi economiche, epidemie, tendenze e mode, la diffusione di notizie e opinioni. Parallelamente, i dati d'interesse possono trovarsi nelle forme più eterogenee: segnali GPS, ma anche dati su fenomeni meteorologici, sulla storia clinica dei pazienti di un ospedale, transazioni al supermercato, eccetera.

In particolare, ogni fenomeno che possa essere descritto da dati diventa analizzabile a livello microscopico. Ogni aspetto dell'attività umana può essere studiato e possono trovarsi, a seconda dell'obiettivo di analisi: schemi ricorrenti; istanze che non si comportano come previsto (*outlier detection*); regole associative che descrivono il comportamento degli utenti e la sequenza prevista delle loro azioni; raggruppamenti latenti in base a una determinata dimensione sulla base dei quali fare profilazione degli utenti; eccetera.

2.1 L'altra faccia dei Big Data: i rischi per la privacy

Pedreschi definisce il diritto alla conoscenza un "bene comune", e in quanto tale necessario e prezioso per chiunque, in modo diretto o indiretto. Proteggere questo bene comune e accertarsi che la sua fruizione avvenga nel rispetto di tutti gli attori coinvolti è un'esigenza che consegue spontaneamente.

L'enorme disponibilità di ogni sorta di informazione è una prospettiva allettante e apre innumerevoli scenari di ricerca; ad esempio, predire con largo anticipo la diffusione di una malattia epidemica permette di prioritizzare gli interventi sanitari preventivi nelle aree più bisognose. Pur avendo benefici incontestabili, però, si presenta la possibilità, seppur remota, di risalire alle fonti di questa conoscenza e ai *dati personali* di singoli, specifici individui. Con "dato personale" si intende qualsiasi informazione che identifica o rende identificabile una persona fisica. Generalmente si distingue tra *dati identificativi* (dati anagrafici, fotografie, ...) e *dati sensibili*, che possono rivelare "l'origine razziale ed etnica, le convinzioni religiose, filosofiche o di altro

genere, le opinioni politiche, l'adesione a partiti, sindacati, associazioni od organizzazioni a carattere religioso, filosofico, politico o sindacale, lo stato di salute e la vita sessuale" [6].

È molto comune che un dataset includa, nella sua forma originale, informazioni che sono riconducibili a specifici individui - un nome, una data di nascita, un indirizzo IP. Ciò diventa potenzialmente pericoloso nel momento in cui si dispone di una chiave per collegare dati su uno stesso individuo provenienti da fonti diverse - informazioni sanitarie, preferenze e opinioni espresse sui social network, dati collegati alle spese online e ai metodi di pagamento. Spesso, il modo in cui tali informazioni vengono registrate non è trasparente (anche se il recente provvedimento del Garante della privacy [8] ha molto migliorato il quadro nazionale, come discusso nel paragrafo 3.1.1), e l'utente medio non sempre è consapevole di quanti e quali dati sensibili stia diffondendo con le sue operazioni quotidiane.

Come puntualizza Pedreschi, siamo tutti fruitori e allo stesso tempo produttori di Big Data, per cui la divulgazione di conoscenza con riguardo alla protezione della sfera privata è, al contempo, prerequisito e obiettivo della Social Data Science, che ha riportato l'umanità al centro dell'investigazione scientifica.

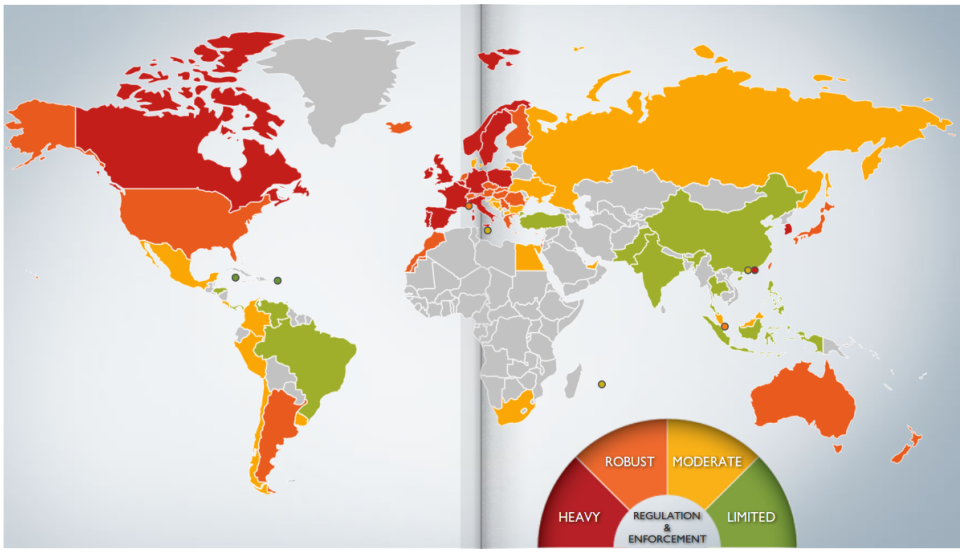

Figura 2: Una panoramica delle diverse leggi in atto nel mondo per la protezione dei dati personali; colori più caldi indicano normative più definite in materia, mentre colori all'altro estremo indicano un corpo legislativo appena accennato. [Fonte: `http://dlapiperdataprotection.com/\#handbook/world-map-section/c1_BR/c2_SG`]

3 La normativa europea

Come visibile a colpo d'occhio in Fig. 2, il quadro europeo in termini di interesse e impegno per la protezione dei dati sensibili è abbastanza incoraggiante, dato che la legislazione in materia è corposa.

L'Unione Europea stabilisce che i dati personali possano essere raccolti legalmente sotto specifiche condizioni e per motivi legittimi. Stabilisce inoltre che le organizzazioni che aggregano dati personali debbano impegnarsi a proteggere gli stessi da utilizzi inappropriati. In particolare, l'UE si preoccupa di assicurare la massima protezione di dati sensibili in tutti i Paesi membri, ovvero di contrastare leggi nazionali che abbassino gli standard di protezione promulgati a livello europeo.

Il 25 gennaio 2012 la Commissione Europea propose una riforma estensiva[1] della normativa sulla protezione dei dati sensibili contenuta nella Direttiva del Parlamento Europeo del 24 ottobre 1995[2], il cosiddetto **pacchetto protezione dati**. Questa riformulazione si è resa necessaria per due ragioni: in primo luogo, la normativa allora vigente era inadeguata a regolamentare il diritto alla privacy sul Web e per le attività online; in secondo luogo, le leggi nazionali dei singoli Paesi, nella pressante esigenza di dotarsi di una legislazione in materia, avevano proceduto in direzioni diverse, rendendo ancora più urgente una riformulazione guidata dall'UE.

Il fondamento di tale riassetto rimane l'art. 8 della Carta dei diritti fondamentali dell'Unione Europea[3]. L'approvazione del pacchetto protezione dati richiede l'intervento sia del Parlamento europeo, sia del Consiglio UE, che al momento hanno proceduto a una prima lettura. L'iter dovrebbe concludersi nei primi mesi del 2016.

Il pacchetto protezione dati si avvale di due diversi strumenti:

- una proposta di Regolamento, che discute "la tutela delle persone fisiche con riguardo al trattamento dei dati personali e la libera circolazione di tali dati" e andrà a sostituire la Direttiva 95/46;

- una proposta di Direttiva che concerne la "regolamentazione dei settori di prevenzione, contrasto e repressione dei crimini, nonché all'esecuzione delle sanzioni penali, che sostituirà (ed integrerà) la decisione qua-

[1] Il comunicato stampa è disponibile al seguente URL: `http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm`

[2] Contrassegnata dalla sigla 95/46/CE, pubblicata nella GUCE L. 281 del 23.11.1995 (p. 31).

[3] L'articolo recita: "Ogni individuo ha diritto alla protezione dei dati di carattere personale che lo riguardano. Tali dati devono essere trattati secondo il principio di lealtà, per finalità determinate e in base al consenso della persona interessata o a un altro fondamento legittimo previsto dalla legge. Ogni individuo ha il diritto di accedere ai dati raccolti che lo riguardano e di ottenerne la rettifica. Il rispetto di tali regole è soggetto al controllo di un'autorità indipendente."

dro 977/2008/CE sulla protezione dei dati personali scambiati dalle autorità di polizia e giustizia". [7].

In preparazione alla stesura di questi strumenti, la Commissione Europea aveva promosso indagini statistiche nei Paesi membri allo scopo di saggiare l'opinione pubblica in materia di privacy sul web. In Italia tale studio fu realizzato tra il novembre e il dicembre 2010 [2]; oltre a domande sull'utilizzo generale della rete, il sondaggio chiedeva agli intervistati di esprimere i propri sentimenti (fastidio, indifferenza, apprezzamento) rispetto a comuni strategie di web marketing, come ad esempio l'inserimento di annunci ad hoc su determinati siti in base alle ricerche compiute, verso cui gli italiani risultarono significativamente meno ostili (43%) rispetto alla media europea (54%), così anche come verso la pratica di fornire informazioni personali al fine di ottenere servizi gratuiti (56% d'accordo contro il 29% della media europea). L'indagine fece emergere una tendenza generale degli italiani ad avere maggior fiducia, rispetto ai colleghi in Europa, che il trattamento dei loro dati fosse lecito e che ci fossero mezzi adeguati per il controllo dei loro dati personali sul web. Per una parte del campione, il dato insolito potrebbe anche spiegarsi con una comprensione solo parziale della potenziale minaccia alla privacy.

Recentemente è stata realizzata un'indagine simile su un campione di 28.000 cittadini, nota al pubblico sotto il nome di "Eurobarometro". I risultati sono stati divulgati nel giugno 2015 [3]; il dato principale, coerente con i sondaggi passati, riporta che i cittadini europei lamentano la mancanza di controllo sui propri dati (67%) e non ripongono fiducia nei venditori online (62%). Ben il 70% degli intervistati teme che i propri dati siano utilizzati per scopi diversi da quelli dichiarati, e addirittura l'89% trova essenziale che la regolamentazione in materia sia comune a tutti i Paesi.

Il 30 giugno la Commissione ha ribadito che il completamento di tale riforma è un obiettivo prioritario, specialmente nell'ottica di rendere più snelli e proficui - sia per i cittadini, sia per le imprese - i processi della *digital economy*, ossia tutti i processi economici supportati dalla tecnologia e, in particolare, realizzati sul web [10] [13]. Si stima che il valore dei dati sensibili dei cittadini europei sarà mille miliardi di euro nel 2020 [11], per cui è di vitale importanza fornire misure di protezione adeguate.

In conclusione, il pacchetto protezione dati promuove un uso corretto dei Big Data, il cui utilizzo da parte delle 100 principali imprese produttrici nell'UE, si stima, potrebbe far risparmiare fino a 425 miliardi di euro. Al tempo stesso, il rispetto della privacy dei cittadini produrrà un "circolo virtuoso tra la protezione di un diritto fondamentale, la fiducia del consumatore e la crescita economica" [11, p. 4]

3.1 La normativa sulla privacy in Italia

In Italia è in vigore il "Codice in materia di protezione dei dati personali", emanato con il d.lgs. 30 giugno 2003, n. 196, che implementa le direttive 95/46/CE and 2002/58/CE del Parlamento Europeo. Tale Codice è stato emendato in seguito alla direttiva 2009/12/CE con il d.lgs. 69/2012 per regolare la raccolta e il trattamento di dati di traffico e geolocalizzazione.

Una volta che la Commissione Europea approverà il regolamento sulla protezione dei dati personali proposto nel 2012, questo sostituirà il d. lgs. 196/2003.

La proposta ruota intorno a diversi punti cardine, tra cui il generale *principio di trasparenza*: l'informazione destinata all'*interessato* (termine con cui ci si riferisce all'utente) deve essere semplice da capire e accessibile (nel senso del termine che afferisce al dominio della *accessibilità sul Web*). In secondo luogo, la proposta affronta la questione della *portabilità del dato*, ovvero la possibilità di trasferire i propri dati da una piattaforma - ad esempio un social network - ad un'altra, e quella del *diritto all'oblio*, che regola le modalità secondo cui si può richiedere la rimozione completa dei propri dati da uno specifico servizio; quest'ultimo punto deve, naturalmente, adattarsi agli obblighi di legge. Infine, gli utenti hanno diritto a non essere sottoposti a una profilazione automatica in base alle loro attività sul web.

Sottotende a questa proposta il cosiddetto *Privacy by design principle*, che si potrebbe forse tradurre come "principio della protezione a priori della privacy"; esso si propone di risolvere l'annosa questione della privacy nel data mining, perlomeno per la maggior parte delle applicazioni. Il principio, molto semplice, intima di tener conto delle possibili minacce alla privacy fin dagli stadi iniziali di tutti i procedimenti che utilizzano dati (a qualsiasi titolo: divulgazione, analisi, trasformazione, sintesi). Si procede quindi verso una protezione della privacy *by default*, cioè con misure preventive, anche a costo di un peggioramento della qualità dei dati di partenza o della conoscenza estratta che è, per la maggior parte delle applicazioni, del tutto trascurabile [15], [19].

Il Codice stabilisce che i dati di traffico devono essere eliminati o resi anonimi quando non più necessari alle comunicazioni elettroniche; tuttavia, l'utilizzo prolungato fino a 6 mesi può essere negozionato con le parti in causa, che hanno sempre la facoltà di revocare la licenza d'uso[4]. Quanto ai dati di geolocalizzazione, questi possono essere utilizzati in forma anonima o conseguentemente al consenso dell'utente, che può sempre essere revocato.

[4] `http://dlapiperdataprotection.com/#handbook/online-privacy-section/c1_IT/c2_SG`

3.1.1 La normativa sui cookie

I *cookie* sono informazioni registrate dal browser durante la navigazione in specifici siti web, che riportano alcuni dati della sessione in corso, quali ad esempio il nome e l'indirizzo del server, identificatori, ora degli accessi, e così via. La persistenza dei cookie può variare e può essere direttamente stabilita dall'utente. L'utilizzo tipico è velocizzare l'autenticazione ai siti web che la richiedano e raccogliere informazioni sugli utenti e sulla loro attività sul servizio web, così come sulle loro preferenze di navigazione (si pensi ad esempio alla scelta della lingua e della valuta mostrate in un sito di e-commerce) [5].

Al di là di questo servizio, i cookie possono anche essere utilizzati per l'*user profiling*, ovvero la profilazione degli utenti intesa come monitoraggio delle loro attività, abitudini e/o desideri di acquisto, allo scopo di assegnare, per esempio, un certo utente a un determinato profilo di utenti che hanno interessi comuni. Ciò accade perché il terminale con cui si accede a Internet viene anch'esso identificato, e perciò può venire utilizzato, insieme al resto delle informazioni contenute nei cookie, in siti web diversi da quelli abitualmente consultati. Una volta che il riconoscimento è avvenuto, il contenuto della pagina, i pop-up e le pubblicità vengono adattate in base al profilo costruito.

Il carattere stesso dei cosiddetti "cookie di profilazione" li rende un potenziale pericolo per la privacy degli utenti. La normativa europea dispone che l'utente debba sempre essere consapevole e favorevole alla raccolta di informazioni su cookie. Con il provvedimento dell'8 maggio 2014 [9], il Garante ha disposto che ogni sito web che utilizza cookie debba chiaramente dichiarare ad inizio sessione: a) che fa uso di cookie di profilazione a scopi pubblicitari; b) se quello è il caso, che i cookie possono essere inviati a terze parti[5]; c) un collegamento URL a un'informativa più dettagliata per la consultazione; d) che, proseguendo con la navigazione, si acconsente all'uso dei cookie.

4 Il trattamento della privacy in Data Mining

L'espressione *data mining* denota una fase del processo di *Knowledge Discovery in Database* (KDD, cfr. Fig. 3, [16]) che consiste nello scoprire pattern interessanti in grandi dataset, tramite l'adozione di tecniche e strumenti mutuati dal mondo dell'intelligenza artificiale, del machine learning e della statistica. Il data mining è un processo essenziale in ogni ambito scientifico, che viene utilizzato sia per ricercare conoscenza latente nei dati, sia

[5] È anche possibile disattivarne completamente l'uso tramite le impostazioni di ogni browser moderno. Un'altra possibilità è attivare la navigazione anonima, che però non registra l'attività avvenuta sul web neppure nella cronologia.

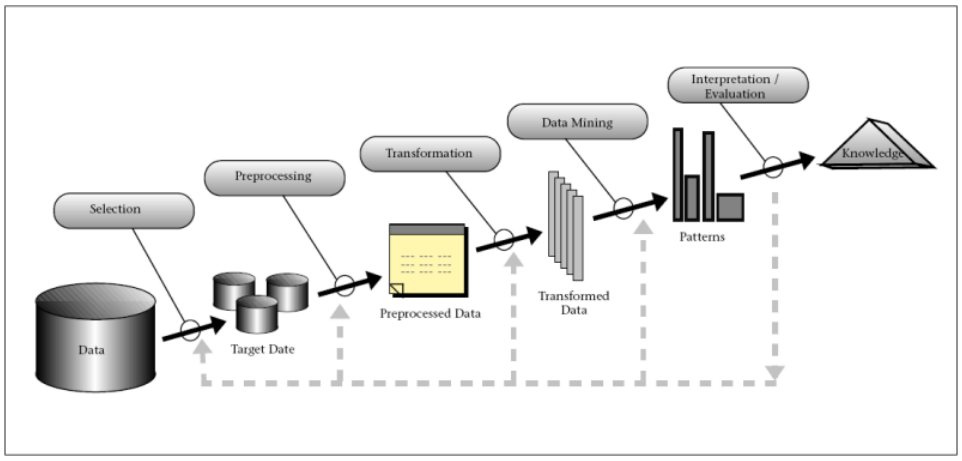

Figura 3: Le fasi del processo di Knolwedge Data Discovery [16].

in approcci semiautomatici o automatici per tentare di scoprire pattern e fenomeni d'interesse in grandi quantità di dati.

Come già discusso, attraverso l'analisi di quantità massicce di dati si possono scoprire regole, modelli, sequenze ricorrenti in dataset diversi e provenienti da fonti diverse. Ogni aspetto dell'attività umana, dai sistemi di trasporti al risparmio energetico, può beneficiare della conoscenza derivata dall'analisi di questi dati tramite metodi di data mining.

Le tecniche di data mining si applicano ai più svariati ambiti: dal business (analisi delle transazioni dei clienti e ricerca di pattern ricorrenti allo scopo di pianificare azioni di marketing), alla biologia e alla genetica (ricerca di anomalie), all'analisi dei social network (identificazione dei nodi più importanti di una rete), e così via. Il vantaggio che si può ottenere da queste applicazioni è notevole, ma molti dataset contengono informazioni sensibili il cui utilizzo solleva annosi problemi di privacy; per tale ragione è parte stessa del data mining l'individuazione, il trattamento (anonimizzazione, rimozione, trasformazione) dell'informazione sensibile. Si parla quindi oggi di **Privacy Preserving Data Mining**, cioè del compito di produrre validi modelli e pattern a partire dai dati, senza divulgare informazioni private [15]. La definizione di "informazioni private", naturalmente, viene regolamentata da Paese a Paese.

5 Approcci al Privacy Preserving Data Mining

Come già accennato, il problema della protezione della privacy non è omogeneo, né si può stabilire a priori quali informazioni si debbano proteggere dalla divulgazione senza tener presente il tipo di dataset in input e il tipo di analisi richiesta.

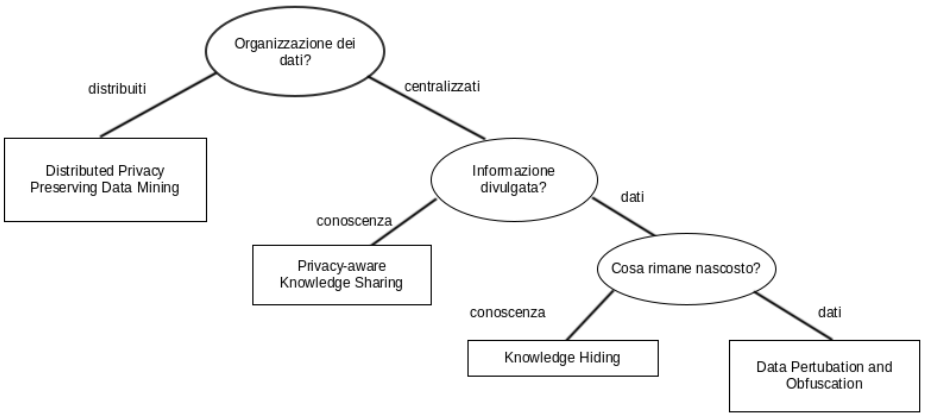

Figura 4: Approcci al Privacy Preserving Data Mining.

In Fig. 4[6] si presenta una panoramica di quattro approcci fondamentali alla protezione della privacy nel Data Mining, così definiti:

1. *Distributed Privacy Preserving Data Mining.* Questo termine racchiude tutte le metodologie in cui il dataset originario viene partizionato e le partizioni vengono distribuite su fonti diverse. Il task di data mining avviene localmente per ogni dataset distribuito, e infine i risultati vengono combinati, facendo sì che nessuna parte in causa abbia accesso a dati contenuti in altre partizioni. Si veda il paragrafo 5.1 per un approfondimento sulle tecniche di computazione distribuita.

2. *Knowledge Hiding.* Si tratta di diffondere i dati in input opportunamente modificati, poiché è noto che determinate analisi possono portare alla luce regole, pattern o modelli (quindi conoscenza) che dovrebbero invece rimanere segreti. L'operazione di modifica, in questo caso, è nota come *data sanization.*

3. *Privacy-aware Knowledge Sharing.* Si tratta di condividere la conoscenza risultata dall'analisi, avendo cura invece di nascondere i dati originari. Il quesito etico cruciale in questo gruppo di algoritmi è se il fatto stesso di sottoporre i dati ad un certo task d'analisi costituisca o no una violazione della privacy.

4. *Data Perturbation and Obfuscation.* Lo scopo di questa famiglia di approcci è proteggere la privacy individuale. I dati in input vengono, anche in questo caso, divulgati dopo modifiche ad hoc, così da rendere impossibile il recupero dei record originali, che tradirebbero

[6]Schema adattato e tradotto in Italiano dal materiale didattico del corso di Data Mining tenuto dai docenti Pedreschi e Giannotti presso l'Università di Pisa, reperibile all'indirizzo `http://didawiki.di.unipi.it/doku.php/dm/start`.

la privacy dei cittadini. L'ambiziosa sfida di ogni approccio che preveda la perturbazione dei dati è mantenere una **distribuzione dei dati** (anche limitata ad alcuni attributi d'interesse) tale che sia ancora possibile estrarre pattern, regole e modelli significativi (*distribution reconstruction*).

I primi due approcci, per loro natura, riguardano la privacy di aziende, corporazioni e, in generale, entità collettive che, poiché vincolati dal segreto industriale o da altri obblighi di riservatezza, necessitano di proteggere la privacy degli utenti nei database, tipicamente dei clienti.

Gli ultimi due approcci, invece, sono orientati alla protezione delle informazioni sensibili del singolo individuo; i dati sono sottoposti ad anonimizzazione (*k-anonimity*, *l-diversity*, v. sottoparagrafi 5.2, 5.3, 5.4) o randomizzazione (v. sottoparagrafo 5.5), con la difficile sfida di mantenerli utili a specifici task d'analisi, come l'estrazione di regole associative. In questa famiglia di approcci i risultati dell'analisi vengono pubblicati senza alcuna alterazione.

Le famiglie di approcci che utilizzano tecniche di perturbazione o anonimizzazione affrontano un'ulteriore sfida, ovvero quella di proteggere i dati a un costo computazionalmente accettabile; infatti, i dataset reali hanno decine, talvolta centinaia di attributi (in letteratura, tale problema è comunemente noto come *curse of dimensionality*), e ciò rende molto difficile la manipolazione e la trasformazione sistematica dei dati. Nelle prossime sezioni sono analizzati in dettaglio alcuni problemi e algoritmi specifici nel campo della protezione della privacy.

5.1 Metodi crittografici distribuiti

Come appena visto, negli approcci distribuiti l'informazione di interesse è debitamente partizionata e localizzata, in parti di per sé insufficienti, in diversi repertori di dati. Il set di attributi rimane lo stesso in ogni distribuzione se il dataset in input viene "tagliato" orizzontalmente, mentre è sempre incompleto, naturalmente, nel caso di un taglio verticale. È possibile interrogare le diverse fonti di dati tramite l'uso di protocolli crittografici che ricompongono l'informazione richiesta senza rivelare alcun dato confidenziale.

Tali metodi sono utilizzati in tutti quei casi in cui gruppi diversi condividono uno scopo, cioè un'esigenza di analisi, e desiderino condividere solo dati aggregati, senza cedere i dati originari di cui sono proprietari.

5.2 K-anonimity

Alcuni attributi dei dati possono essere degli pseudo identificatori se usati congiuntamente, cioè possono identificare univocamente dei record (per esempio zip code e data di nascita). L'idea alla base di un algoritmo di k-anonimity è di diminuire la specificità dell'informazione descritta da un

	Età	Sesso	CAP	Diagnosi
1	18	M	56123	Positivo
2	23	F	56126	Negativo
3	18	M	56124	Negativo
4	23	F	56125	Positivo
...	...	...	...	...

Tabella 1: Ogni paziente è univocamente identificabile, perché ogni record individua uno e un solo individuo.

	Età	Sesso	CAP	Diagnosi
1	18-25	M	5612*	Positivo
2	18-25	F	5612*	Negativo
3	18-25	M	5612*	Negativo
4	18-25	F	5612*	Positivo
...	...	...	...	...

Tabella 2: Il ricorso a informazioni più generiche per l'età e la provenienza è sufficiente ad anonimizzare tutti i record in esame.

dato *record* (cioè l'insieme di attributi-valori riferiti ad un'istanza dei dati), così che esso non possa più essere distinto da almeno altri k - 1 record.

Si considerino i record in Tabella 5.1, riferiti a immaginari pazienti di un ospedale esaminati per la presenza del virus HIV. Si noti che ogni paziente è univocamente identificabile, perché la combinazione dei suoi dati (età, sesso, CAP) individua uno e un solo individuo. Si immagini una situazione in cui tali dati, seppur privi di identificativi quali nome e cognome o codice fiscale, siano confrontati con il registro dei pazienti, che contiene ancora i dati anagrafici; non è difficile immaginare uno scenario in cui il paziente 1 possa essere inequivocabilmente identificato, così come gli altri. Attributi come CAP (o *zip code*) e data di nascita sono chiamati *quasi-identifiers* [19], in quanto rendono esponenzialmente più semplice l'identificazione di un record anonimizzato.

Nel caso dei dati di mobilità, l'informazione spazio-temporale è, a sua volta, un quasi-identifier. Si pensi a una persona che tutti i giorni lascia la località A per andare a lavoro nella località B alla stessa ora, e viceversa si sposta da B ad A ogni sera alla stessa ora. Questa informazione così granulare sarebbe già di per sé sufficiente a identificare percorsi individuali distinti. Se queste informazioni si potessero consultare i registri telefonici, sarebbe possibile collegare le telefonate (che si agganciano a una specifica cella, e sono quindi distinguibili per posizione spaziale) alle traiettorie, e quindi alle persone [21].

Si consideri ora la Tabella 5.1, in cui si è scelto di operare intervalli consecutivi per la descrizione dell'attributo *Età* e di ricorrere al più generico

	Età	Sesso	CAP	Diagnosi
1	18-25	M	5612*	Positivo
2	18-25	M	5612*	Positivo
3	18-25	M	5612*	Positivo
4	18-25	F	5612*	Positivo
...	...	...	...	...

Tabella 3: Il problema della L-diversity: impedire che informazioni sensibili su un singolo individuo vengano estratte dall'osservazione del gruppo di record che condivide gli stessi valori.

codice di avviamento postale di città, 5612*, invece dei CAP 56121..56128. Con questi semplici accorgimenti, tutti i record dell'esempio in esame sono anonimizzati. Un set di record che ha gli stessi valori per gli attributi che sono quasi-identifiers è chiamato *classe di equivalenza* [17].

In conclusione, i dati sensibili nel dataset in input possono essere protetti con la seguente semplice euristica: se ogni quasi-identifier nel dataset compare $\geq k$ volte, allora ogni record della tabella deve essere "nascosto" in $\geq k$ record; ciò equivale a dire che ogni individuo viene "mimetizzato" da $\geq k$ suoi *peers*, cioè da altri individui col medesimo quasi-identifier nei dati originari. Un'altra via percorribile è la rimozione degli attributi che sono quasi-identifiers nelle fasi di pulizia e preparazione dei dati, ma non sempre essi sono superflui per l'analisi da svolgere come nell'esempio in Tabelle 1 e 2.

5.3 L-diversity

Il presupposto della l-diversity è opposto alla tecnica precedente. Si prenda il caso in Tabella 5.2, in cui i record 1-3 condividono non solo gli stessi valori per gli attributi quasi-identifier, ma *anche* gli stessi attributi contenenti dati sensibili (ad esempio diagnosi o storia medica del paziente). In questo caso, si può immaginare uno scenario in cui la diagnosi (positiva) di un paziente che si sa essere registrato nel database può essere facilmente recuperata, in quanto *tutti* i pazienti di sesso maschile, di età compresa tra i 18 e i 25 anni e residenti nell'area 5612* sono stati diagnosticati positivamente. In altre parole, l'omogeneità del gruppo di record che condivide gli stessi valori per certi attributi può mettere a rischio la privacy dell'intero gruppo.

La soluzione a questo problema è ricostruire i gruppi e intervenire sui valori dei quasi-identifier in modo che ci siano almeno l valori diversi per l'attributo target (in questo caso *Diagnosi*) [17].

5.4 T-closeness

In scenari ancora più complessi, la parte in causa interessata all'informazione privata contenuta nel database potrebbe avere informazione a priori sulla distribuzione dei valori di un attributo target (per esempio ricorrendo a dataset esterni che è possibile collegare al dataset oggetto dell'attacco). Il principio di *t-closeness* è una misura in realtà molto semplice: la distribuzione dei valori dell'attributo "sensibile" in ogni gruppo di quasi-identifier deve essere vicina alla distribuzione dei valori di quell'attributo nell'intera tabella [17]; in questo modo, l'eventuale informazione pregressa a disposizione dell'hacker non è più informativa.

5.5 Randomizzazione

L'idea alla base della randomizzazione è di manipolare i dati da divulgare, distorcendone la natura quanto basta per eliminare qualsiasi collegamento troppo esplicito tra il dato sensibile e l'individuo (cruciale specialmente nel caso degli EHR, *Electronic Health Records*).

Il punto di forza di questo metodo è che può essere eseguito nella fase di raccolta dati, in quanto non richiede conoscenza a priori sulla distribuzione degli stessi. Esiste però anche un'altra faccia della medaglia: la **densità locale** dei record non viene tenuta in considerazione, quindi record che descrivono fenomeni *outlier* sono trattati come gli altri [19, p.36]; Kargupta e il suo gruppo di ricerca [18] hanno dimostrato che esistono casi in cui il dataset originale può essere ricostruito.

5.5.1 Noise insertion

Un particolare tipo di randomizzazione consiste nell'aggiungere rumore ai dati, così che i dati originali non possano più essere ricostruiti. Si definisca l'insieme dei dati originali come $X = x_1...x_m$; si ricavi un *set di rumore* denotato da $N = n_1...n_m$. Il nuovo dataset $Z = z_1...z_m$ si ottiene aggiungendo n_i a ogni record $x_i \in X$.

Quanto pesante sia tale manipolazione è stabilito da un compromesso tra il livello di granularità (e accuratezza) della conoscenza da rilasciare e l'impatto sulla privacy dei soggetti.

La noise insertion è spesso affiancata da altre tecniche di anonimizzazione, come la rimozione di attributi "parlanti" e di quasi-identifiers.

5.5.2 Differential privacy

Questo approccio risponde ai casi in cui il dataset non viene direttamente rilasciato né modificato, ma può essere interrogato da terze parti. Data la query di un utente, i risultati ricevuti si basano sul dataset originario, ma sono alterati in una misura determinata dal modello di differential privacy,

pur restando ancora utili allo scopo della ricerca. Il vantaggio di questo approccio è che il dataset non viene mai completamente rilasciato o reso disponibile in modalità diverse da quanto desiderato; tuttavia, è necessario un continuo monitoraggio delle query sottomesse per prevenire episodi di inferenze o collegamento a dati reali.

6 Conclusioni

Il pacchetto protezione dati ha una portata che va ben oltre la protezione della privacy dei cittadini europei: garantire che i dati personali siano protetti, anonimizzati e non riconducibili ai singoli individui permette il *libero utilizzo* degli stessi, aprendo la strada a innumerevoli task di analisi e quindi, virtualmente, a un vantaggio competitivo inestimabile derivato dalla conoscenza estratta da questi dati. Il settore pubblico è poi il palcoscenico ideale per applicazioni di data mining e KDD, in quanto serve un pubblico vastissimo e offre applicazioni che portano un immediato benificio a tutti gli attori coinvolti.

La direttiva 95/46/CE del Parlamento europeo e del Consiglio del 24 ottobre 1995 [1] ha formato un organo, noto come *Article 29 Data Protection Working Party*, che ha il compito di tutelare i cittadini in materia di elaborazione dei dati personali e circolazione dei suddetti.

Dal lavoro di tale organo, che vede al suo interno un rappresentante della Commissione Europea e rappresentanti di ogni Paese facente parte dell'UE, è stato prodotto il documento "Opinion 05/2014" [14]. Esso offre un quadro dei vantaggi e dei limiti delle tecniche di data mining più utilizzate per la protezione dei dati personali - brevemente presentate nel paragrafo precedente - e promuove l'utilizzo combinato di più tecniche, da decidersi caso per caso. Inoltre, il documento raccomanda ulteriori misure da intraprendersi nel caso persista un rischio di identificazione (problematica che riguarda tutte le tecniche citate) [4].

La disposizione stabilisce che debba essere impedito a tutte le parti di:

- individuare univocamente un individuo nel dataset;
- (essere in grado di) collegare due record in uno stesso dataset (si pensi a rapporti di parentela precedentemente noti a chi escogita l'attacco) o in dataset diversi;
- inferire (con una probabilità significativa) il valore di un attributo dal valore di altri attributi.

Si noti che se un hacker riesce a stabilire che due individui appartengono a un certo gruppo, ma non riesce a identificarli, allora la tecnica utilizzata ha garantito il primo requisito, ma non il secondo.

Sfortunatamente, nessuna delle tecniche citate può rispondere a questi requisiti con piene garanzie. Il margine di rischio residuo è tale che il più volte citato privacy by design principle acquisisce un ruolo ancor più cruciale: la chiave alla protezione della privacy risiede in una pianificazione attenta e controllata dei processi e degli strumenti coinvolti nell'iter di estrazione di conoscenza; tutte le tecniche adottate in un secondo momento per anonimizzare o randomizzare i dati o l'informazione da essi estratta devono piuttosto essere concepiti come ulteriori misure di sicurezza e robustezza del framework, e non come principale strumento di protezione.

Assodato che i nostri dati personali non sono, ancora, completamente al sicuro, risulta altresì fondamentale individuare in una fase iniziale qualsiasi rischio potenziale connesso al rilascio di determinate informazioni.

Riferimenti bibliografici

[1] *Direttiva 95/46/CE del Parlamento europeo e del Consiglio, del 24 ottobre 1995, relativa alla tutela delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati.* Official Journal of the European Communities. I(281), pagine 31-50, 24 ottobre 1995. Reperibile online: `http://194.242.234.211/documents/10160/10704/Direttiva+95+46+CE.pdf`.

[2] *Atteggiamento nei confronti della protezione dei dati e dell'identità elettronica nell'Unione europea*, 2012. `http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_fact_it_it.pdf`.

[3] *Data protection Eurobarometer. Factsheet*, 2012. `http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_eurobarometer_240615_en.pdf`.

[4] *Article 29 Working Party issues Opinion on Anonymisation techniques*, 2015. `http://www.epsiplatform.eu/content/article-29-working-party-issues-opinion-anonymisation-techniques-0#sthash.gsf24gDB.2EwrdRwd.dpuf`.

[5] *Cookie e privacy: dalla parte degli utenti [4020961]*, 2015. `http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/4020961`.

[6] *Cosa intendiamo per dati personali*, 2015. `http://www.garanteprivacy.it/web/guest/home/diritti/cosa-intendiamo-per-dati-personali`.

[7] *Il nuovo "pacchetto protezione dati": Proposta di Regolamento generale sulla protezione dei dati personali e Proposta di Direttiva sulla protezione dei dati personali nelle attività di contrasto.*,

2015. http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/4443361.

[8] *Individuazione delle modalità semplificate per l'informativa e l'acquisizione del consenso per l'uso dei cookie (pubblicato su GU n. 126 del 3-6-2014).*, 2015. http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/3118884.

[9] *Individuazione delle modalità semplificate per l'informativa e l'acquisizione del consenso per l'uso dei cookie - 8 maggio 2014 [3118884]*, 2015. http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/3118884.

[10] *Protection of Personal Data*, 2015. http://ec.europa.eu/justice/data-protection/index_en.htm.

[11] *The EU Data Protection Reform and Big Data, Factsheet*, 2015. http://ec.europa.eu/justice/data-protection/files/data-protection-big-data_factsheet_web_en.pdf.

[12] *Welcome to the yotta world*, 2015. http://www.economist.com/node/21537922.

[13] *What will the EU Data Protection Reform bring for startup companies and Big Data?*, 2015. http://ec.europa.eu/justice/newsroom/data-protection/news/150415_en.htm.

[14] *WP216 Opinion 05 2014 on "Anonymisation Techniques onto the web*, 2015. http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/3070528.

[15] Aggarwal, C.C. e P.S. Yu: *Privacy-Preserving Data Mining: Models and Algorithms*. Advances in Database Systems. Springer US, 2008, ISBN 9780387709925. https://books.google.com.sg/books?id=8Vr3PtZ3Y7wC.

[16] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases*. AI magazine, 1996. pagina 37.

[17] Fosca Giannotti, Anna Monreale, Dino Pedreschi: *Mobility Data and Privacy*. Nel *Mobility Data Modeling, Management, and Understanding*. Springer, 2013. pagine 174–193.

[18] Kargupta, Hillol, Souptik Datta, Qi Wang e Krishnamoorthy Sivakumar: *On the privacy preserving properties of random data perturbation techniques*. Nel *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pagine 99–106. IEEE, 2003. pagine 99–106.

[19] Monreale, Anna: *Privacy by Design in Data Mining*. tesi di dottorato, Dipartimento di Informatica, Università di Pisa, 2011.

[20] Pedreschi, Dino: *Social Data Science*. presentato al Seminario di Cultura digitale, disponibile a `http://www.labcd.unipi.it/seminari/dino-pedreschi-social-data-science/`.

[21] Sushil Jajodia, Steve Noel, Brian O'Berry: *Topological analysis of network attack vulnerability*. Nel Vipin Kumar, Jaideep Srivastava e Aleksandar Lazarevic (curatori): *Managing Cyber Threats: Issues, Approaches and Challenges*. Springer, 2005. pagine 248–266.

Tutti gli indirizzi web riportati sono stati visitati nel dicembre 2015.