La de-anonimizzazione dei dati personali. Il caso del dataset Netflix

> *"Releasing the data and just removing the names does nothing for privacy. If you know their name and a few records, then you can identify that person in the other (private) database" (Vitaly Shmatikov).*

> *"A ticking privacy time bomb" (Marc Rotenberg, Electronic Privacy Information Center).*

Introduzione

Nel 2006 Netflix aveva pubblicato una classifica di 10 milioni di film, da parte di 500.000 clienti, come parte di una sfida rivolta agli sviluppatori e agli appassionati di *machine learning* di trovare sistemi di raccomandazione migliori di quello che l'azienda stava utilizzando in quel torno di tempo. Come base di dati, su cui applicare gli algoritmi, aveva fornito delle variabili anonimizzate, ovvero aveva rimosso i dati personali e sostituito i nomi con numeri casuali, al fine di proteggere la privacy degli utenti. Ciò era stato attentamente vagliato e nel rilascio dei dati l'azienda affermava testualmente, in una domanda retorica, pubblicata nella FAQ di accompagnamento del rilascio pubblico dei dati: "*Is there any customer information in the dataset that should be kept private?*". La risposta che l'azienda dava era la seguente. "*No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy. […] Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and*

that data was subject to perturbation. Of course, since you know all your own ratings that really isn't a privacy problem is it?".

In questo breve testo si illustra come due ricercatori dell'Università del Texas, Arvind Narayanan e Vitaly Shmatikov, siano riusciti a de-anonimizzare i dati di Netflix confrontando classifiche dei film (*ratings*) e marcatori temporali (*timestamp*) con informazioni pubblicamente disponibili nell'Internet Movie Database (IMDb). Benché siano passati diversi anni la rilevanza della tematica è tuttora di cogente attualità, a maggior ragione dopo l'approvazione del GDPR comunitario e del crescente interesse che viene attualmente annesso a tale tematica. La loro ricerca illustra alcuni problemi di sicurezza intrinseci con i dati anonimizzati e, a parere di chi scrive, è ancora più rilevante che tale tematica sia stata sollevata nel lontano 2008, più di dieci anni fa. Detto in altri termini, come è stato possibile che con un piccolo numero di dati, relativi a un determinato individuo, nessuno dei quali, presi a sé stante, erano univocamente identificabili, inseriti in un contesto più ampio, si siano trasformati, di fatto, in un vero e proprio identificativo personale?

La de-anonimizzazione del *dataset* Netflix

La metodologia che hanno seguito i due studiosi non è stata quella di fare un *reverse engineering* dell'intero dataset quanto piuttosto quello di incrociare i dati rilasciati pubblicamente da Netflix con quegli utenti, scelti in maniera casuale, i quali avevano inserito alcune recensioni di film, con i loro nomi, nell'IMDb. Sebbene i *records* di IMDb siano pubblici, la scansione del sito per ottenerli è contro i termini di servizio di IMDb, quindi i ricercatori hanno utilizzato solo alcuni campioni degli stessi (un piccolo campione di 50 utenti IMDb) per dimostrare la validità del loro ragionamento e del loro algoritmo.

La domanda di ricerca era dimostrare quante poche informazioni siano veramente necessarie per de-anonimizzare le informazioni presenti nei datasets rilasciati pubblicamente. "*We investigate the following question: How much does the adversary need to know about a Netflix subscriber in order to identify her record if it is present in the dataset, and thus learn her complete movie viewing history?*" (Narayanan & Shmatikov, 2008:8).

I due ricercatori che hanno lavorano sui dati anonimizzati da Netflix non hanno cercato di capire minuziosamente le identità delle persone – come altri avevano fatto in precedenza, ad esempio, con il *database* di ricerca di AOL. Essi hanno semplicemente comparato i dati anonimizzati con un sottoinsieme già identificato di dati simili presenti sul sito IMDb: ovvero, una tecnica di *data mining* standard la quale può essere facilmente replicata. La loro è stata una semplice assunzione di fondo. "*We expect that for Netflix subscribers who use IMDb, there is a strong correlation between their private Netflix ratings and their public IMDb ratings. […] In many cases, even a handful of movies that are rated by a subscriber in both services would be sufficient to identify his or her record in the Netflix Prize dataset*" (Idem, 2008:12).

La tematica è di interesse generale in quanto, a solo titolo esemplificativo, Alphabet Inc., con il *database* relativo alle ricerche compiute dagli utenti sul suo motore *Google Search*, potrebbe facilmente de-anonimizzare una base dati pubblica di acquisti sulla rete oppure concentrarsi su ricerche di termini medici per de-anonimizzare un *dataset* sparso relativo alle cartelle cliniche personali. I siti di commercio elettronico, i quali mantengono informazioni dettagliate sui clienti e sugli acquisti da loro compiuti in rete, potrebbero utilizzare i propri dati per de-anonimizzare parzialmente i dati di qualsiasi altro *database* di grandi dimensioni, seppur rilasciati in forma anonima. Ulteriore esempio, potrebbe

essere quello di una società che acquista una base di dati da altre società commerciali e potrebbe essere in grado di de-anonimizzare la maggior parte dei *records* ivi contenuti.

Ciò che i ricercatori dell'Università del Texas sono riusciti a dimostrare è, allora, che questo processo non è così impossibile da replicare e, in diversi casi, non richiede nemmeno una mole elevata di dati. Se si eliminano, difatti, i 100 migliori film che tutti guardano, ovvero i risultati più comuni, il resto dei comportamenti individuali in rete costituisce un'impronta digitale (*fingerprint*), in molti casi, facilmente riconoscibile. "*Let us summarize what our algorithm achieves. Given a user's public IMDb ratings, which the user posted voluntarily to reveal some of his movie likes and dislikes, we discover all ratings that he entered privately into the Netflix system*" (Idem, 2008:13).

E ciò non è solo vero per la visione dei film e delle serie televisive compulsate su Netflix, lo si ribadisce ancora una volta, ma vale anche per le abitudini, spiccatamente idiosincratiche, riguardanti la lettura di libri, gli acquisti su internet, le ricerche svolte sui motori di ricerca, ecc.. Il risultato duraturo della ricerca dei due ricercatori è di aver dimostrato che le tecniche di anonimizzazione applicate a *datasets* sparsi non rappresentano una reale protezione dei dati degli utenti. "*Sanitization techniques from the k-anonymity literature such as generalization and suppression do not provide meaningful privacy guarantees, and in many case fail on high-dimensional data. Furthermore, for most records simply knowing which columns are non-null reveals as much information as knowing the specific values of these columns. Therefore, any technique such as generalization and suppression which leaves sensitive attributes untouched does not help*" (Idem, 2008:14).

Gli abiti (*habits*) digitali hanno né più e né meno la stessa valenza degli abiti tradizionali: rappresentano dei comportamenti stabili dell'essere umano, trasferiti

attualmente in un contesto digitale, e perciò essi stessi passibili di accumulazione, aggregazione e incrocio da parte di svariate base di dati. Come è stato autorevolmente affermato dagli stessi autori, "*in contexts such as behavioral advertising, scholars argue that the power to influence behavior is deeply problematic even if the data is never linked to a real-world identity*" (Narayanan & Shmatikov, 2019).

Le contromisure per questa situazione potrebbero essere, purtroppo, inadeguate. Netflix avrebbe potuto randomizzare il suo set di dati rimuovendo un sottoinsieme di dati, modificando i marcatori temporali (*timestamp*) oppure aggiungendo errori deliberati nei numeri ID univoci utilizzati per sostituire i nomi. I due ricercatori hanno però dimostrato che tutto ciò serve solo a rendere il problema leggermente più difficile. L'algoritmo di de- anonimizzazione di Narayanan e Shmatikov si è dimostrato, difatti, sorprendentemente robusto e funzionante anche con dati parziali, dati randomizzati ed anche con dati volutamente erronei. "*It works even if the published records have been perturbed, if only a subset of the original dataset has been published, and if there are mistakes in the adversary's background knowledge*" (Narayanan & Shmatikov, 2008:2). Con solo otto valutazioni date ai film visionati e con delle date che potrebbero avere un errore fino a due settimane, l'algoritmo è riuscito a identificare in modo univoco il 99% dei *records* nel set di dati. "*Our conclusion is that very little auxiliary information is needed for de-anonymize an average subscriber record from the Netflix Prize dataset. With 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, 99% of records can be uniquely identified in the dataset. For 68%, two ratings and dates (with a 3-day error) are sufficient*" (Idem, 2008:11).

Dopodiché, tutto ciò di cui hanno avuto bisogno i due ricercatori sono stati un po' di dati passibili di

identificazione: può essere IMDb come qui, nel caso di specie, ma potrebbe essere anche un *blog*, qualsiasi altro *social networks*, oppure una base di dati differente. La conclusione è che potrebbe bastare anche un piccolo database *ad hoc*, il quale contenga dei dati personali, affinché si possa de-anonimizzare un *dataset* sparso molto più grande.

In precedenza, altre ricerche erano giunte alla medesima conclusione. Utilizzando dati anonimizzati, pubblici, relativi al censimento del 1990, Latanya Sweeney aveva scoperto che l'87% della popolazione negli Stati Uniti, 216 milioni su 248 milioni, potrebbe venire identificata in modo univoco solo grazie al codice postale a cinque cifre, combinato con il sesso e la data di nascita. Circa la metà della popolazione degli Stati Uniti è probabilmente identificabile per sesso, data di nascita e città, paese o comune in cui risiede la persona. Se si volesse espandere l'ambito geografico a un'intera contea la percentuale si ridurrebbe del 18% ma rimarrebbe ancora significativa. In un lavoro che rappresenta la ripetizione di tale esperimento, compiuto nel 2006, sui dati censuari del 2000, Philippe Golle riporta che "*our results generally agree with the findings of* [Sweeney, 2000], *although we find that disclosing one's gender, ZIP code and full date of birth allows for unique identification of fewer individuals (63% of the US population) than reported in* [Sweeney, 2000]". Dal suo breve studio emerge che già solo la data di nascita è in grado di ordinare le persone in migliaia di strati diversi e che riesce a disambiguare in maniera efficace le persone che hanno effettuato il censimento.

Tutto ciò ha profonde implicazioni per il rilascio pubblico di dati anonimi. Da un lato, i dati anonimizzati costituiscono la base del lavoro quotidiano di migliaia di ricercatori, pubblici e privati. In generale, sarebbe da apprezzare che moltissime società, e persino la pubblica amministrazione rilasci i dati in forme pubblicamente disponibili (*open data*). Nel caso precedentemente citato, ovvero quello di AOL, il CTO

della società avrebbe sì compiuto un'azione non disdicevole nel rilasciare pubblicamente il *database* di dati anonimizzati per motivi di ricerca, però la conseguenza finale è stata l'esposizione pubblica delle ricerche personali compiute dagli utenti, un danno reputazionale enorme subito da AOL e, finanche, le dimissioni dello stesso CTO, insieme all'intero team di ricerca, dopo la protesta sollevata dal giornalista del New York Times. Solo per fare un altro esempio, i dataset medici anonimizzati sono estremamente preziosi per gli avanzamenti scientifici della società intera, ovvero per studi di farmacologia su larga scala, studi di *follow-up* a lungo termine e così via ma inevitabilmente espongono i nomi dei partecipanti ad essere de-anonimizzati con una relativa facilità, con le inevitabili implicazioni di privacy che si possono facilmente immaginare. E che l'opinione pubblica attuale mostra di non voler più accettare, nemmeno per ragioni di benessere collettivo, come l'insufficiente diffusione dell'*app* Immuni dimostra assai bene, anche nel caso italiano.

Conclusioni

Nel lavoro seminale di Sweeney (2002) la de-anonimizzazione era basata sulle variabili demografiche in quanto nei *databases* di allora erano questi i dati predominanti di cui si voleva preservare l'anonimità. Tuttavia, oggigiorno sono cresciute, a dismisura, le modalità di raccolta dei dati ed essi sono disponibili in formati digitali, sparsi nella variegata geografia dei siti *web*. Ebbene in questi casi sono possibili tecniche di *data mining* per de-anonimizzarli grazie alla presenza di alcuni attributi di granularità maggiore come le recensioni dei film su IMDb o la cronologia di navigazione dei *browers* oppure i luoghi digitali visitati i quali, tutti, possono essere utilizzati a questo fine nonostante il set di dati rilasciato pubblicamente contenga centinaia o migliaia di osservazioni individuali. E ciò si è dimostrato vero per campi quali i *social networks*, i dati genetici, i dati di geolocalizzazione, i dati relativi alle carte di credito, la

cronologia di navigazione dei *browsers*, lo stile di scrittura, il codice sorgente del software, il codice compilato del software, ecc.. "*This line of research has firmly established that high-dimensional data is inherently vulnerable to de-anonymization. This is also supported by theoretical evidence. When we consider the fact that 33 bits of entropy are sufficient to identify an individual uniquely among the world's population, these research findings should be no surprise*" (Narayanan & Shmatikov, 2019). Il contributo duraturo di Narayanan e Shmatikov, nel loro paper del 2008, e di molti altri ricercatori che li hanno seguiti su questo percorso di ricerca, è stato dunque quello di mettere in luce come nonostante, e forse proprio per questo, una grandissima mole di dati (*big data*) disponibili sulla rete digitale sono inerentemente vulnerabili alla de-anonimizzazione. "*A small number of data points about an individual, none of which are uniquely identifying, are collectively equivalent to an identifier*" (Idem, 2019).

In conclusione, nell'era del capitalismo della sorveglianza (Zuboff, 2019), in cui tutte le entità, sia statuali sia commerciali, raccolgono dati su tutte le persone e per tutto il tempo nei quali sono in rete, l'anonimizzazione dei dati personali e sensibili può risultare una pratica operativa molto fragile e rischiosa più di quanto si potesse immaginare al tempo in cui essa era stata proposta e conseguentemente ampiamente adottata, sia nel settore governativo e, forse ancor più, in quello privato.

In questa situazione, appare assai opportuno chiedersi il "Che fare?" di leniniana memoria. Un ricercatore come Bruce Schneier sostiene che i sistemi di anonimizzazione non dovrebbero essere resi operativi e divulgati se prima non dimostrano di aver superato una fase preparatoria di tecniche di *adversarial attacks*. Egli fa un paragone con i sistemi di crittografia, di cui è un esperto, e altre modalità di rilascio del software i quali vengono previamente messi a

disposizione della comunità *open source* per essere testati ai fini di sicurezza. Solo quelle modalità di anonimizzazione che dovessero aver superato questi controlli incrociati dovrebbero essere poi sviluppati, messi in produzione e resi pubblicamente disponibili. Allo stesso modo, così come è opinione comune che è non prudente implementare un sistema crittografico prima che venga rigorosamente attaccato, ebbene, si dovrebbe fare la stessa cosa con l'anonimizzazione dei dati personali. E, come ogni altra modalità nel campo della sicurezza, l'anonimizzazione è sempre un difficile compromesso. Ci sono vantaggi e ci sono rischi corrispondenti all'utilizzo o meno di determinate tecniche e metodologie. Ciò implica, evidentemente, l'emergenza di nuovi paradigmi regolativi, a livello normativo, di cui l'attuale GDPR comunitario può essere forse considerato solo il primo esempio di una lunghissima serie di altri di prossima emanazione.

Referenze bibliografiche

- **Golle P.** (2006), *Revisiting the Uniqueness of Simple Demographics in the US Population,* WPES '06,October 30, pp. 4.
- **Narayanan A. & Shmatikov V.** (2008), *Robust de-anonymization of large sparse datasets,* 29th IEEE Symposium on Security and Privacy, pp. 111-125.
- **Narayanan A. & Shmatikov V.** (2009), *De-anonymizing social networks*, 30th IEEE Symposium on Security and Privacy, pp. 173-187.
- **Narayanan A. & Shmatikov V.** (2019), *Robust de-anonymization of large sparse datasets. A decade later*, Manuscript, May 21, pp. 5.
- **Sweeney L.** (2000), *Uniqueness of Simple Demographics in the U.S. Population*, Carnegie Mellon University, Laboratory for International Data Privacy, LIDAP WP4, pp. 34.
- **Sweeney L.** (2002), *k-anonymity. A model for protecting privacy*, International Journal of Uncertainty,

Fuzziness and Knowledge-Based Systems", v.10, n.5, pp. 557-570.
- **Zuboff S.** (2019), *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*, PublicAffairs, New York (NY).

Articolo a cura di **Achille Pierre Paliotta**

Achille Pierre Paliotta, ricercatore INAPP(ex ISFOL), afferente alla Struttura Lavoro e Professioni. Laureato in Sociologia presso l'Università La Sapienza ha conseguito, nel 2015, il Master in *Data Science* presso l'Università di Tor Vergata. Sta attualmente frequentando il Master in *Sicurezza delle informazioni e informazione strategica* presso l'Università La Sapienza. I suoi interessi di ricerca vertono su nuove tecnologie, sicurezza informatica, nuove professioni, contrattazione collettiva, *welfare* occupazionale, *data science*, *blockchain*, *open data* e *deep learning*.

Tra le sue ultime pubblicazioni: *Le politiche innovative di sicurezza nelle città tra tecnologie di riconoscimento e smart cities* ("SINAPPSI", 2020); *Decentralized Bargaining and Measures for Productivity and Occupational Welfare Growth in Italy* ("eJICLS", 2019); *Il sistema della salute in transizione. La cartella sanitaria personale tra commodificazione dei dati e blockchain* ("Salute e Società", FrancoAngeli, 2019); *Nuevas profesiones y técnicas de web data mining en Argentina: el caso del Data Scientist* ("CESCOT", 2018).

https://www.ictsecuritymagazine.com/articoli/la-de-anonimizzazione-dei-dati-personali-il-caso-del-dataset-netflix/