838009
**Andrea Lacavalla**

relatore Marco Ronchi
correlatore Enrico Frumento

# la profilazione nel digital marketing

Best practice e rischi. Progettazione di un sistema di interazione per navigare la digital shadow di un'azienda e dei suoi dipendenti.

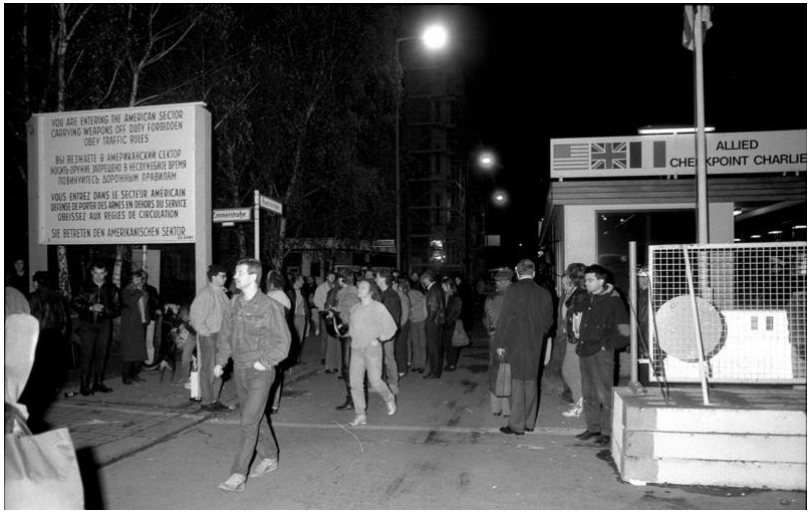
Politecnico di Milano - Scuola del Design
Laurea Magistrale in Design della Comunicazione - A.A. 2015-16

# indice:

# internet e libertà

Il World Wide Web probabilmente non esisterebbe, almeno nei termini in cui lo conosciamo oggi, se la nascita di internet non fosse avvenuta proprio in quel contesto politico che caratterizzò lo scenario mondiale della seconda metà del XX secolo. Non si è trattato semplicemente di specifiche conquiste in campo tecnologico, che hanno dato gli strumenti per collegare tra loro i primi computer. Questo è solo un aspetto tecnico che fa parte di un processo ben più complesso che chiama in causa spinte culturali, storiche e, soprattutto, politiche.

La prima teorizzazione di internet risale a una pubblicazione scientifica degli anni '60. Il progetto fu poi ultimato nel 1969 dalla DARPA, un'agenzia subordinata al Dipartimento della Difesa degli Stati Uniti, come tecnologia militare di spionaggio durante la Guerra Fredda. Nei due decenni successivi, altre nazioni svilupparo-




1989
*Berlinesi attraversano Checkpoint Charlie la sera della caduta del muro*

no la propria rete internet locale, tra cui l'Italia, la Norvegia e l'Inghilterra.
Con il finire della Guerra Fredda si diffuse un nuovo clima politico, e con esso un forte desiderio di apertura e di abbattimento di quelle barriere che si erano consolidate negli anni precedenti: nel 1898 cadde il muro di Berlino e tra il 1992 e 1993 nacque l'Unione Europea.

Fu proprio in un momento come questo che poté nascere il World Wide Web, e lo fece grazie a un fatto talvolta trascurato, ma in realtà cruciale. Non è l'invenzione del protocollo HTTP, prima o poi l'avrebbe inventato qualcun altro. La vera svolta fu quando, riflettendo pienamente quel senso di apertura di fine secolo scorso, nel 1993 il CERN rese pubblica la tecnologia di base di internet. In questo modo veniva permesso a chiunque di implementarla, e di conseguenza questa pubblicazione diede la possibilità agli Stati di lavorare a partire da un'unica base per estenderla in tutto il mondo, anziché costringerli a progettare ciascuno la propria rete da zero. In altre parole, era nato il concetto di open source.
La natura intrinseca del web, come luogo aperto, ha portato la community di early adopter a fare propri valori quali la libertà di espressione e di comunicazione. Ma se questa apertura è l'origine di internet, l'anonimato non può che essere una necessaria conseguenza, per esempio per evitare le conseguenze dell'esprimere opinioni proibite o del pubblicare contenuti senza permesso. E il web permetteva in effetti un buon livello di anonimato: bastava acquistare un dominio per poter pubblicare più o meno qualsiasi cosa senza la necessità di identificarsi.

# il web 2.0 apre a scenari rivoluzionari

Come spesso accade con le nuove tecnologie, in un primo momento il web tendeva a riprodurre elementi già esistenti, limitandosi a veicolarli attraverso la sua inedita struttura; molti siti assomigliavano infatti a delle riviste, in cui periodicamente venivano pubblicati nuovi contenuti di notizie, riflessioni, fotografie. In pochi anni tuttavia si capì che un nuovo tipo di interazione era possibile: permettere a coloro, che prima erano chiamati visitatori, di caricare contenuti facendoli così diventare utenti. Questa novità, dal mio punto di vista rivoluzionaria, tanto quanto la decisione di far diventare la tecnologia di internet open source, è considerata così determinante, che oggi è comunemente usato il termine web 2.0 per riferirsi all'uso interattivo dei siti. E, come la rivoluzione precedente, anche questa ha portato con sé delle implicazi-

oni non trascurabili che hanno radicalmente modificato lo scenario d'uso della rete.

Da un punto di vista strettamente pratico, il proprietario del dominio ha più poteri delle altre persone che pubblicano sul suo spazio: per esempio può scegliere che cosa sia possibile pubblicare e cosa no, oppure può ottenere informazioni su chi usa la piattaforma sfruttando gli strumenti di amministrazione o costringendo gli utenti a identificarsi. In questo modo diventa possibile che vengano occasionalmente meno i due principi cardine dell'internet alle sue origini: la libertà di opinione e l'anonimato.
Tuttavia c'era un terzo, forse non altrettanto nobile, principio generalmente associato ad internet, più difficile da far venire meno, e cioè che tutto fosse gratis. Anche se si è trattato più di una supposizione degli utenti che di una reale caratteristica del web, sembrava che avesse perfettamente senso: si paga il consumo della banda, e in cambio si ottengono beni e servizi; inoltre questa associazione mentale era favorita dal fenomeno della pirateria, grazie alla quale con un rischio praticamente nullo ci si poteva liberamente e anonimamente scambiare contenuti che si era soliti pagare, come album musicali, film e videogiochi.
Ma un host non riceve certo i soldi pagati dagli utenti ai provider, quindi come fa a permettersi i server e i domini su cui



caricare le piattaforme? E a pagare un ingegnere perché le costruisca? Inoltre, più sono complesse la struttura e l'interattività a disposizione, più i costi menzionati lievitano. D'altra parte è difficile per una persona accettare di pagare qualcosa che è abituata ad avere gratis. In un momento di tale transizione, era necessario tentare diversi modelli di business per cercare di monetizzare questa nuova, incredibile, fonte di traffico.

# nuovi modelli di business

Il recente dominio di internet ha generato grandi interessi e numerosi scenari possibili per trarre profitto. Dei tanti modelli di business utilizzati oggi, alcuni sono stati copiati da altri servizi già in uso, altri sono stati inventati ex novo, altri ancora sono il frutto dell'elaborazione di vecchi modelli riprogettati secondo convenienza o applicabilità. Data la nat-

ura molteplice e talvolta ibrida delle varianti a disposizione, non è sempre possibile inserire in categorie e schemi fissi le modalità di profitto; tuttavia in genere si considerano esserci circa nove modelli tra i quali è possibile svariare: non è raro infatti che una compagnia utilizzi un modello di business personalizzato, per esempio utilizzandone più di uno contemporaneamente, o un misto tra due diversi.
Queste nove categorie sono generalmente identificate con i loro nomi inglesi.

Il brokerage è uno dei modelli più semplici, che consiste nel convogliare venditori e acquirenti in un'unica piattaforma di compravendita; il profitto è ricavato trattenendo una piccola percentuale sulle transazioni. Per esempio, utilizzano questo modello di business eBay ed Airbnb.

Per i venditori di beni che aggiungono o sostituiscono un punto vendita online a uno offline, il modello è detto di tipo manufacturer: il venditore crea la sua piattaforma di vendita di proprietà, pertanto ottiene del profitto accorciando la catena tra esso e il cliente. Oggigiorno trovare un esempio del genere è facilissimo dal momento che ormai quasi tutti i produttori hanno un proprio punto vendita online, da Apple a Nike, da Mondadori a Ikea.

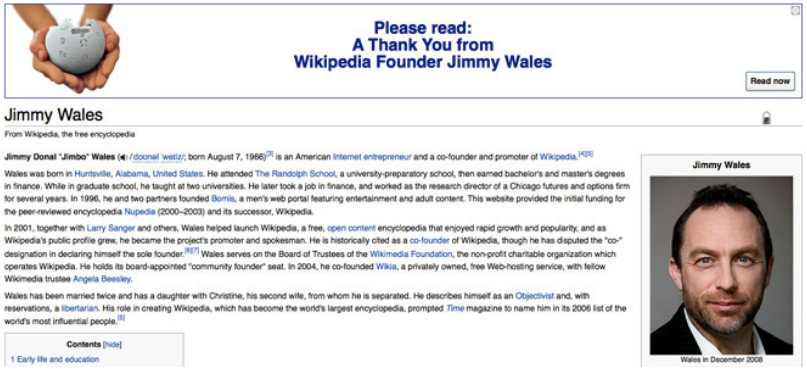
Un modello praticamente identico, detto



merchant, utilizza lo stesso metodo di vendita, e anche qui allo spaccio online può esserne affiancato uno fisico. La differenza sta nel tipo di venditore: in questo caso non rientrano i produttori ma i rivenditori, come Media World o Asos. Naturalmente, questi modelli valgono anche per la vendita di software e altre tipologie di materiali o servizi non necessariamente fisici.

Un modello poco conosciuto è l'affiliate model, in cui un produttore aggiunge un canale di vendita presso un rivenditore online. In genere il produttore ricompensa questo aumento di possibilità di vendita pagando una piccola percentuale delle transazioni al rivenditore. Nel caso in cui l'affiliato non fosse in grado di completare alcuna vendita non rappresenterebbe comunque un costo per il produttore. Utilizza questo modello Amazon, in aggiunta alla vendita diretta di merce (quindi con il modello merchant).

Uno dei modelli che sta trovando grande fortuna recentemente è il subscription based, in cui gli utenti pagano una quota periodica – in genere mensile, ma anche annuale, settimanale o talvolta illimitata - in cambio di poter usufruire di certi servizi e contenuti in stile all you can eat. Due tra i maggiori interpreti di questo modello sono Netflix e Spotify; lo utilizzano moltissimo anche le riviste scientifiche.




2011

*Da anni, il co-fondatore di Wikipedia Jimmy Wales chiede agli utenti di contribuire con una donazione*

Il modello utility propone contenuti on demand, e cioè in cui a ogni azione corrisponde una piccola somma da pagare, ma è decisamente poco utilizzato. Preso in prestito da realtà come luce e gas oppure la telefonia - che fino a qualche tempo fa prevedeva nella maggioranza dei casi un costo al minuto o per SMS - oggi lo si può ritrovare ad esempio in Tuts+.

Fanno parte del modello community tutte quelle realtà che traggono profitto, direttamente o indirettamente, dal contributo della comunità. Alcune modalità sono le donazioni o l'impegno mirato al miglioramento del prodotto con le proprie competenze (Open Source). Un sito che ottiene benefici in entrambi questi modi è Wikipedia.

Uno tra i più comuni è invece il modello di advertising: i contenuti e i servizi vengono messi a disposizione degli utenti, il più delle volte gratuitamente, e vengono adibiti degli spazi alle pubblicità. Proprio



come succede nella televisione, il prezzo è fatto dalla quantità e dalla qualità del traffico generato dai siti ospiti e gli advertiser scelgono il canale di distribuzione in base alla tipologia di utenti prevista. È talmente diffuso che si trovano esempi a ogni livello, da YouTube al peggior sito di streaming pirata.

La nona categoria è un modello molto spesso collegato a quello dell'advertising, e costituisce oggi uno dei metodi più potenti e redditizi tra tutti quelli a disposizione per monetizzare attraverso internet. Parliamo dell'infomediary model, che consiste nel conservare e catalogare i dati lasciati durante la navigazione dagli utenti al fine di rivenderli o utilizzarli per trarre profitto. Questi dati vengono reperiti con modalità diverse e possono essere aggregati al fine di compilare profili univoci degli utenti e delle comunità di cui fanno parte; questo processo è chiamato profilazione, e moltissime compagnie - anche indipendenti - si trovano attualmente in una vera e propria corsa all'oro volta ad ampliare, aumentare e perfezionare questa pratica, ragion per cui ultimamente si stanno facendo degli autentici passi da gigante. E tutto ciò a ragione, poiché le potenzialità paiono enormi.
Essendo, almeno a un primo sguardo, uno dei modelli meno intuitivi, sorgono spontanee alcune domande. Di che tipologia di dati si tratta? Come vengono raccolti

questi dati? Come vengono utilizzati? A chi vengono venduti, e con che fine?

# commerciare i dati: potenzialità e benefici

Scorrendo i modelli di business elencati precedentemente, si può in realtà osservare che la maggior parte di essi è applicabile solamente a quegli scenari in cui è previsto movimento di denaro a prescindere dal medium attraverso cui questo avviene, dal momento che si tratta di attività commerciali o di piattaforme di compravendita. Le uniche vie percorribili per chi non vende materiali o servizi, né ospita movimenti di denaro, sono le ultime tre tipologie: la community, l'advertising e l'infomediary model). Tuttavia scegliere il modello di community non



è semplice; bisogna infatti considerare che è necessario raggiungere un pubblico vasto - o molto facoltoso - e far leva su un sentiment estremamente positivo, oppure un commitment elevatissimo. Può essere il caso di cause sociali o umanitarie, oppure di materiale che ci risolve un grosso problema - da Salvatore Aranzulla che spiega come recuperare dei dati che davamo ormai per persi, a Wikipedia che ci salva dal prendere un 2 a scuola - ma non tutti i siti possono permettersi di contare su questo tipo di introiti per sopravvivere.
Resta fondamentalmente un unico modo di monetizzare il proprio traffico da parte di tutti quei siti che non ospitano compravendite, che sono tra l'altro la netta maggioranza sul web - come il motore di ricerca di Google e i social network, per non parlare dei forum e dei blog: fare advertising. Non c'è quindi da stupirsi che un sistema per migliorare l'efficacia della pubblicità come l'infomediary diventi così popolare, soprattutto se può garantire un differenziale unico rispetto ai media tradizionali quale la possibilità di mostrare annunci mirati.

L'esempio più classico di pubblicità targetizzata è Facebook. Prima di tutto, ha una base di utenti elevatissima, che si avvicina ai 2 miliardi di visitatori unici ogni mese; poi, gli utenti sono tenuti a comunicare il proprio vero nome, la data di nascita e il genere. In più, sono spinti

anche a fornire informazioni circa il proprio luogo di nascita e di residenza, ed è possibile completare il profilo con informazioni anche più specifiche come l'orientamento politico e religioso, le preferenze sessuali, i propri parenti. Bastano queste poche informazioni ad aprire le porte ad annunci molto specifici: per esempio, sponsorizzando una piccola scuola di danza locale a tutte le donne tra i 28 e i 38 anni residenti in paese, sperando che siano mamme di una bambina da iscrivere; il tutto a un prezzo modico naturalmente. Tuttavia si può andare ancora più nel dettaglio, decidendo di mandare l'annuncio a quelle donne della ricerca precedente con "Mi Piace" a pagine legate alla maternità: considerando alla pari le nascite di maschi e femmine, almeno una persona su due di chi vede l'annuncio è un potenziale cliente, una percentuale altissima rispetto ai media tradizionali. Tuttavia, come per esempio è emerso da un'inchiesta di Hannes Grassegger e Mikael Krogerus sulla vittoria di Donald Trump - e come ampiamente dichiarato dalla compagnia stessa che si è occupata di progettare e targetizzare i suoi annunci - sappiamo che è possibile andare ancora più a fondo e indirizzare gli annunci anche a un pubblico raccolto in gruppi di etnia, personalità, indole e profilo psicologico. In questo caso probabilmente a un prezzo meno modico, se non altro per la parcella dei data scientist.



Ma l'uso in ambito pubblicitario non è l'unico: ci sono oggi diversi servizi che possono esistere solo grazie ai dati degli utenti e altri che potrebbero funzionare comunque, ma hanno successo solo grazie all'uso della profilazione. Uno dei casi più interessanti e innovativi tra i servizi che funzionano grazie ai dati è IFTTT, acronimo di if this then that. Si tratta di una piattaforma che coordina diversi social e software per impostare delle azioni automatiche attraverso una sintassi di causa ed effetto, in cui in base a un evento preselezionato su un'app avviene un'azione su un'altra. Per esempio potremmo dargli l'istruzione di mandarci un'email se il prezzo di un'applicazione scende, o di abbassare al minimo la suoneria non appena Waze (app di mappe e navigatore GPS) dovesse rilevare che siamo arrivati al lavoro. Grazie a componenti esterni come Arduino e il GPS dell'automobile, queste azioni possono anche uscire dai nostri dispositivi personali e dare comandi come "quando la mia macchina arriva a casa, apri il garage".
Mentre IFTTT ha bisogno di accedere ad applicazioni, dispositivi e dati esterni, un servizio come Spotify potrebbe tranquillamente funzionare comunicando unicamente con i propri server e senza tracciare le azioni dell'utente. Tuttavia, uno dei suoi punti di forza è proprio la capacità di ricordarsi le azioni dell'utente in merito ai generi preferiti, o gli artisti non graditi, e di indovinare con discreto suc-

cesso i suggerimenti di nuovi brani. Inoltre, l'integrazione con Facebook permette a Spotify di leggere i "Mi Piace" dell'utente e avere così una base da cui partire per scongiurare cattive esperienze nei primi giorni di utilizzo del programma.

Diventa così ben comprensibile perché molte aziende investano così tanto sull'analisi dei dati, sia che cerchino di aumentare i profitti da advertising sia per cercare di proporre servizi innovativi, e perché anche tante compagnie che sono in grado di monetizzare efficacemente in altri modi non vogliano restare fuori dalla corsa, con Amazon in prima fila. Nel prossimo capitolo si analizzeranno gli strumenti con cui questi dati vengono raccolti.

# strumenti e metodi di raccolta dei dati

Nella progettazione di un supporto digitale che usi o generi dati, inclusa una digital strategy, è necessario comprendere il tipo di informazioni con cui si avrà a che fare: non solo dal punto di vista semantico, ma anche degli elementi da cui sono composte. Come vengono veicolate, attraverso che canali e con quali tecnologie, e infine come è possibile elaborarle.

# metadati

Ogni giorno in internet vengono caricati TeraByte su TeraByte di dati. Solo una macchina potrebbe essere in grado di processare una tale quantità di informazioni; tuttavia a un computer mancano le facoltà cognitive di base per poter interpretare una fotografia, un video o un testo, o per ricollegare tra loro elementi che derivano da fonti diverse. Grazie ai progressi della tecnologia, però, oggi esistono alcuni metodi per imitare tali capacità, tipiche dell'essere umano, in modo automatico. Uno di questi è l'interpretazione dei metadati, che sono dei riferimenti associati al file: per esempio una canzone in formato mp3 può contenere non solamente l'audio, ma anche il nome dell'artista, il titolo dell'album, l'anno di uscita e tante altre informazioni.
Mentre il suono ha una natura semantica che può essere interpretata e classificata solo da un essere umano, utilizzando correttamente i metadati è possibile indicizzare e gestire grandi quantità di brani senza doverli ascoltare uno a uno. Si può per esempio isolare e dunque ascoltare un certo album tralasciando gli altri brani, oppure creare una playlist che ripercorra



il cammino artistico di un autore riproducendone l'intera discografia in ordine di tempo.
Alcuni metadati fanno inevitabilmente parte del file, altri vengono scritti automaticamente in fase di generazione o modifica dello stesso, altri ancora possono essere aggiunti manualmente. Per esempio, nel caso di una fotografia, il file porta necessariamente con sé le dimensioni in pixel, ottiene in automatico dalla macchina fotografica le impostazioni dello scatto (come tempo di esposizione e ISO) e può essere arricchito dall'utente con un titolo. Naturalmente non tutti gli apparecchi producono necessariamente gli stessi metadati, allo stesso modo in cui non tutti i fotografi aggiungono le stesse informazioni alle loro foto. In modo simile, è possibile che alcuni metadati non vengano letti da tutti i software. Tuttavia, molti metadati comuni sono generalmente visibili da tutti i principali software, Per esempio la stringa "Autore" di un brano musicale.

Ad ogni modo i metadati non nascono con il digitale, nonostante questo ne abbia esponenzialmente aumentato le possibilità di uso. Per esempio, l'archivio di una biblioteca usava già i metadati, anche se forse in modo meno efficiente di come si può fare oggi. Come i nostri documenti in pdf, le nostre canzoni, le immagini e anche gli stati su Facebook, i libri portano con sé la componente semantica del testo,

ma anche dei riferimenti: nella biblioteca questi sono il nome dell'autore, il numero di pagine, la categoria in cui si trova il libro e la casa editrice, mentre in uno stato su un social ci sono l'ora della pubblicazione, le persone coinvolte nell'attività e la località in cui è stato scritto, oltre ad alcuni dati tecnici che descrivono il dispositivo da cui è stato inviato - se fisso o mobile -, il sistema operativo utilizzato, il browser.
Un buon lavoro di indicizzazione può rendere l'operazione di ricerca di un libro molto più semplice, anche quando partiamo da pochi elementi; e se il database è ricco, dettagliato e coordinato con altri dati esterni sarà più semplice anche ottenere informazioni sui testi addirittura prima di leggerli. Per esempio, se sappiamo che un autore scrive romanzi brevi e lunghi saggi storici, ci basterà sapere il titolo e il numero di pagine di una sua opera per poterne prevedere approssimativamente il contenuto; a maggior ragione se questo è scritto in collaborazione con un altro autore, poiché incrociando la varietà di tematiche trattate da entrambi, si riduce considerevolmente lo spettro degli argomenti disponibili.

In modo simile, i metadati di ciò che pubblichiamo possono rendere molto più semplice un lavoro di catalogazione automatica. Per esempio, scrivere uno stato su Facebook che recita "Stasera birrificio!" è immediatamente comprensibile



a un essere umano, specialmente se ne conosce l'autore. Si potrebbe sapere che questo abita nella zona di una famosa birreria, e si può ipotizzare che andrà con un gruppo di amici, probabilmente quelli che frequenta di solito. Al contrario, per una macchina queste basilari abduzioni sono impossibili; tuttavia, se questo post viene arricchito con dei metadati, anche una macchina può giungere a conclusioni simili: basta taggare le persone con cui si passerà la serata e fare il check-in nella birreria grazie al GPS del telefono.

Sembrerebbe dunque che produciamo metadati solo quando creiamo o pubblichiamo dei contenuti. E in effetti è così, ma col tempo è stato realizzato che anche la nostra semplice navigazione in un sito genera contenuti, anche quando non carichiamo nulla in modo diretto: cliccare


*I metadati di una fotografia scattata da uno smartphone*

su un link, andare alla seconda pagina di un articolo, fare una ricerca con dei filtri sono tutte azioni che inviano una richiesta specifica ai server del sito, i quali hanno bisogno di leggerla per dare la risposta che vuole l'utente. Tutte queste richieste vengono trascritte dai server all'interno di cosiddetti log files, ovverosia dei verbali con tutta la comunicazione tra server e utente, i quali portano ovviamente con sé dei metadati. Se indicizzati bene possono aiutare a catalogare queste interazioni, per esempio aggregando le azioni per argomenti più cliccati, o per orari di maggiore attività, generando dei report quantitativi sull'attività degli utenti che aiutino un aggiustamento o una nuova progettazione dell'interazione.

## user experience

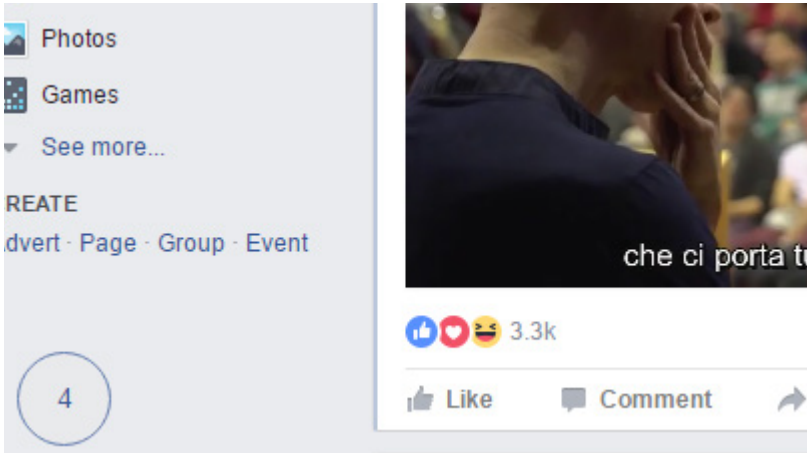
I metadati possono quindi aggiungere dettagli ai contenuti creati, o al massimo informazioni quantitative circa il loro utilizzo, per esempio quando in iTunes vediamo quante volte è stato riprodotto



un brano. Tuttavia, nella maggioranza dei casi non sono in grado di fornire indicazioni sul modo in cui se ne usufruisce.
I servizi più evoluti in fatto di raccolta di dati, che sono molto interessati alla qualità di consumo degli stessi, hanno dunque sviluppato delle tecnologie che permettono loro di leggere l'esperienza dell'utente in rapporto a ciò che visita e agli elementi con cui interagisce. Uno di questi servizi è certamente Facebook, il cui modello di business, come abbiamo visto, consiste in una fortissima targetizzazione dei contenuti sponsorizzati: chi paga per un'inserzione può infatti scegliere con notevole precisione il tipo di pubblico a cui rivolgersi, e la capacità di Facebook di raggiungerlo è fondamentale perché il suo mercato cresca e mantenga la sua posizione di élite. Per questo motivo, i metadati delle azioni degli utenti non bastano più: è vero che sapere data di nascita, luogo di provenienza, sesso e argomenti di interesse (questi attraverso i "Mi piace") è importantissimo per una promozione targetizzata, ma è anche possibile che qualcuno abbia mentito circa le proprie generalità, o non abbia messo dei "Mi piace" a pagine a cui non vuole essere associato dai suoi amici, o ancora che non abbia reagito (con una reaction, una condivisione o un commento) a un elemento solo perché pubblicato da un contatto poco simpatico, o con cui ci si sentirebbe a disagio a stabilire un'interazione.
Per cercare di avere una visione meno vi-

ziata da certe dinamiche, sono state sviluppate alcune tecniche di monitoring che vanno oltre i log file, che pure consentono un primo livello di analisi del comportamento dell'utente.

Una delle più efficaci è il page tagging, ovverosia una tecnica complementare ai log file che permette di indicizzare aree o elementi di una pagina (come paragrafi, immagini, pulsanti) e di tracciare l'interazione dell'utente con questi, per esempio calcolando il tempo di mouse hover su un oggetto o il tempo in cui ci si sofferma su una certa area durante lo scroll, il tutto senza dover mandare alcuna richiesta ai server.
Sfruttando sempre il caso studio di Facebook, come dichiarato nei loro stessi Termini del Servizio, essi sono in grado di raccogliere dati dall'analisi dell'attività sulla piattaforma, in particolare, i post




*L'add-on Data Selfie usa il page tagging per contare quanto ci soffermiamo su ogni post di Facebook*



che seguiamo di più, le persone sulle cui attività ci soffermiamo più spesso, le tematiche che più facilmente ci fanno fermare durante lo scorrimento.
Si instaura in questo modo una specie di circolo virtuoso, in cui viene analizzata l'esperienza utente con (anche) l'obiettivo di sviluppare nuove funzionalità per migliorarla, renderla più piacevole e completa e, di conseguenza, avere ancora più elementi da studiare in futuro. Questo significa da un lato, proporre nuove interazioni, e dall'altro, migliorare l'integrazione delle pubblicità e affinare le tecniche di raccolta dei dati per massimizzare gli introiti. Non è un caso che Facebook e i software che possiede siano sempre orientati verso l'ampliamento dei dati raccolti e dell'integrazione tra i servizi, di cui un esempio recente è stata la controversa associazione degli account con i profili Whatsapp.

In realtà, la maggior parte degli elementi che si aggiungono alla ricchezza dell'esperienza d'uso di una pagina è anche in grado di fornire dettagli in più circa l'uso della stessa, e gli esempi sono numerosi. Un altro caso è quello delle API, acronimo di Application Programming Interface: si tratta di fatto di script che con poche linee di codice sono in grado di integrare parte di un servizio all'interno di un altro, o di trasmettere certi dati all'esterno senza per questo far accedere l'esterno ai propri database. Un esempio abbastanza

comune è quello dell'integrazione di YouTube in Facebook; quando condividiamo un video del tubo sul nostro profilo non dobbiamo effettivamente fare l'upload del video, né questo viene rielaborato da Facebook in alcun modo, tant'è vero che il lettore resta quello di YouTube, inclusa la barra di navigazione e i pulsanti del volume e di impostazioni. Questo significa che Facebook sfrutta un blocco di codice preparato appositamente dai programmatori di YouTube che permette di aprire un canale di comunicazione attraverso cui dal primo leggiamo dati del secondo, senza per questo cambiare sito o avere accesso completo all'archivio di YouTube.
Un altro esempio sono le API per fare data crawling: nonostante di recente alcuni servizi le abbiano rimosse, tra cui Facebook stesso e LinkedIn, sono molti quelli che permettono ancora di fare crawling sui propri utenti, tra questi Twitter. Nonostante la user interface possa sembrare limitata da questo punto di vista, non consentendoci di aggiungere tutti i filtri che vorremmo alle nostre ricerche, è possibile utilizzare le loro API ufficiali (pubbliche e gratuite) per impostare ricerche personalizzate e salvare i risultati come tabelle indicizzate. È possibile farlo anche su altri servizi, naturalmente a patto che i post e gli elementi presi in considerazione abbiano visibilità pubblica. In caso contrario, l'API è impostata per non essere in grado di vedere i dati.



# cookie e 3rd party tracker

Abbiamo visto che sia l'aumento delle possibilità di generare dati personali, e sia l'incremento della capacità di interpretarli hanno portato a una maggior diffusione in rete dell'uso da parte delle aziende della profilazione.
Chi si è accorto subito delle potenzialità di questa tendenza e ha deciso di investirci si è trovato un vero e proprio pozzo di petrolio sotto i piedi. A beneficiarne sono stati soprattutto quei servizi che, per funzionare, facevano già un uso massiccio e sapiente dei dati di navigazione. Tuttavia non tutti sono in grado di produrre da sé dei database di questo genere, in certi casi per mancanza di risorse tecniche ed economiche, in altri perché il bacino di utenza non è abbastanza ampio da poter costituire una base statistica e numerica sufficiente. Un esempio potrebbe essere quello di un quotidiano online locale: per quanto possa essere interessato a migliorare il suo servizio di pubblicità, sfruttando i dati ricavati dalla profilazione, potrebbe non avere non avere né gli strumenti adeguati a elaborarli, né

un numero di utenti sufficiente a generarli. Esigenze come questa da parte di chi non riesce a istituire un proprio sistema di tracciamento hanno contribuito alla nascita dei tracker di terze parti, ovverosia aziende specializzate nella gestione di grandi quantità di dati che offrono questo servizio alla clientela.

Per riuscire a registrare informazioni sui siti clienti, queste compagnie utilizzano i cookie. Si tratta di semplici file di testo con all'interno alcune istruzioni che vengono lette dai server che si visitano. Queste istruzioni vengono aggiornate a ogni visita per tenere traccia di alcuni aspetti della navigazione; per esempio possono tenere memoria di alcune preferenze di visualizzazione di una pagina, o di alcune impostazioni salvate in una visita precedente. E naturalmente vengono registrati anche i dati di navigazione mediante log files e page tagging: link visitati, movimenti all'interno della pagina e tutte quelle altre attività specificate nei capitoli scorsi. Durante la visita successiva il server legge il cookie e mantiene tutte le preferenze delle volte precedenti.
Di recente l'Unione Europea ha emanato delle direttive sulla gestione dei cookie a cui generalmente ci si riferisce con l'espressione "Cookie Law", anche se di fatto le specifiche leggi - stilate sulla traccia dell'UE - sono scritte e applicate localmente e non coinvolgono i soli cookie. I tracker di terze parti infatti possono

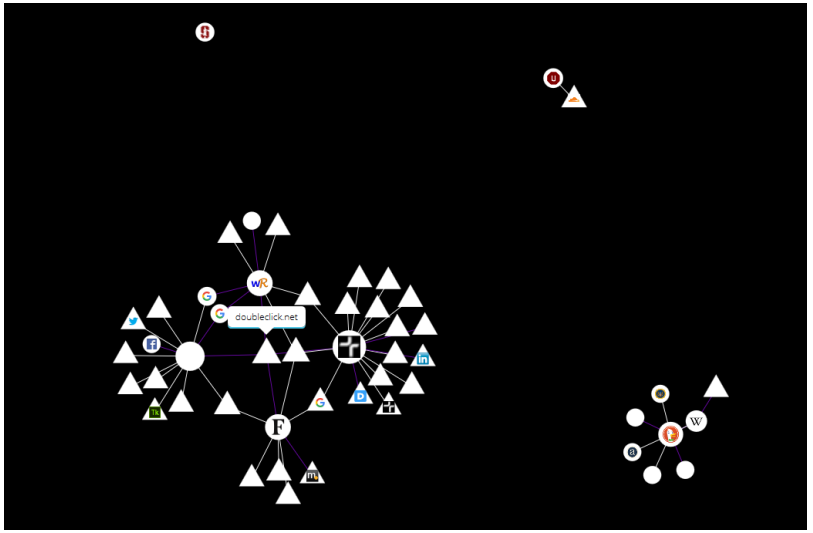


viaggiare anche attraverso altri mezzi tecnici, come la memoria locale di Flash e HTML5, anche se si tratta, per diverse ragioni, di realtà minori e perciò di non primaria importanza per questa legge. Tra gli argomenti a sfavore di queste, ci sono il sempre più scarso utilizzo di Flash, la grande agilità dei cookie (leggerissimi e universalmente compatibili) e questioni di sicurezza: i file di testo non sono in grado di portare virus, al contrario di contenuti Flash e link.

Quello che fanno le compagnie è dunque progettare dei cookie ad hoc per tenere traccia di tutte quelle informazioni che ritengono siano utili ai fini del tracciamento e del tipo di ricerca che si vorrà fare a partire da esse; successivamente sviluppano un software che, attraverso degli algoritmi, sarà in grado di indicizzare, riordinare e dare significato a quei dati, dopo averli inseriti in un database.
A seconda del tipo di servizio acquistato sarà possibile progettare algoritmi più o meno complessi, oppure utilizzarne alcuni generici ritenuti validi come standard; sarà inoltre possibile scegliere se lasciar fare tutto in automatico alla macchina oppure adoperarsi per rintracciare degli insight all'interno dei dati.
Tuttavia, nonostante la presenza dei tracker di terze parti sia ormai la normalità, la conoscenza di questo ecosistema è ancora molto limitata.

Una panoramica del loro funzionamento si può ottenere attraverso quegli strumenti in grado di rilevare i tracker. Come approfondiremo meglio più avanti, Mozilla è una delle software house che hanno più a cuore la privacy online, e infatti, è sua una delle estensioni più note in grado di svolgere questo compito. Si tratta di Lightbeam, un servizio che, in una scheda a parte del browser, illustra con una mappa intuitiva e aggiornata in tempo reale, sia i siti mano a mano visitati e sia i relativi tracker che sono entrati in funzione, collegandoli tra loro. Così facendo si può individuare subito se un componente di terze parti ha tracciato la nostra navigazione in più siti diversi; i tracker più importanti, o che accidentalmente abbiamo incontrato un gran numero di volte, andranno a formare dei

*La mappa di Lightbeam mostra lo scambio di informazioni tra i siti e i tracker attivi*






veri e propri nodi all'interno della topologia di rete attorno ai quali orbitano i siti. Altre volte capiterà che i siti più grossi siano essi stessi dei nodi, dal momento che intorno a loro ruotano numerosi tracker. Non è raro infine trovare che un sito da noi visitato sia in realtà a sua volta un tracker di terze parti su un altro.

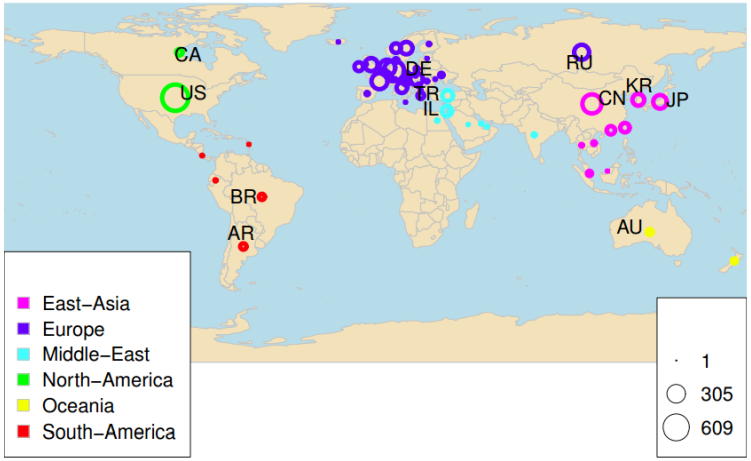
Ho utilizzato l'add-on, con il fine di analizzarne il comportamento, a più riprese e per diversi intervalli di tempo: la mappa che si è venuta gradualmente a comporre è curiosa.
In base a una prima prova di utilizzo, ho notato due macrogruppi centrali più grandi, per dimensioni e numero di collegamenti tra parti interne, ma del tutto scollegati tra loro. Si trattava di Google e Facebook. Una notevole quantità di tracker direttamente collegati erano in realtà delle entità separate sul grafico, ma evidentemente di reciproca proprietà; tante volte è possibile determinare ciò dall'URL del tracker, che rimanda per esempio a googleadservices.com oppure a fbcdn.net, che, come favicon, porta la stessa "f" di www.facebook.com. C'erano poi altri snodi abbastanza importanti, per esempio il Guardian, il cui sito appariva notevolmente tracciato.
Dopo qualche giorno ho rilevato che i gruppi con più collegamenti tendevano ad accorparsi tra loro, mentre i macrogruppi principali continuavano a rimanere isolati. Questo scenario bizzarro e del tutto

imprevedibile è continuato finché non ho visitato You Tube: solo dopo aver fatto accesso al sito, anche i macrogruppi si sono connessi tra loro.
È stato pertanto interessante rilevare che i due colossi – Google e Facebook – tendono a evitarsi. Ciò ha fatto sorgere in me il sospetto che possa esistere una sorta di policy - richiesta da almeno uno dei due - che vieti a un 3rd party tracker di fornire i suoi servizi anche al rivale. Ci stiamo però muovendo nel campo delle ipotesi e delle congetture.
Ad ogni modo non sorprende che fosse proprio YouTube l'anello mancante tra i due, dal momento che è di proprietà di Google, ma è ben integrato con Facebook e anzi ne porta all'interno delle API (il pulsante "Condividi su FB") e viceversa, Facebook ha integrato il player – già precedentemente citato – di You Tube.
Dopo ulteriori giorni di utilizzo, la mappa ha iniziato a diventare sempre più articolata e complessa, oltre che a risultare via via più pesante sulla RAM: la tendenza degli elementi è stata quella di uscire dalla schermata, e navigandola non è stato raro riscontrare problemi di lag.
Ho quindi supposto che, a questo punto della navigazione, difficilmente sarebbero rimasti isolati dal gruppo centrale alcuni nodi. Tuttavia, dando uno sguardo più attento alla mappa, ho scoperto che alcuni siti continuavano a non collegarsi ai macrogruppi. Si tratta perlopiù di siti molto vecchi, probabilmente non aggior-



nati da anni, oppure accademici, che presumibilmente non puntano a fare profitto né innovazione della UX e dunque non necessitano di alcun tipo di servizio di tracciamento.
Un'ultima considerazione riguarda i siti che dicono esplicitamente di non tracciare, come Mozilla e Duckduckgo. Va riconosciuto che, pur essendo due tra i siti che ho visitato più spesso - l'uno per ragioni di studio su questa stessa tesi, l'altro per essere il motore di ricerca di mia scelta e predefinito del browser - sono stati dei piccoli satelliti indipendenti che orbitavano al di fuori del nucleo del grafico per buona parte del tempo, e sono stati circondati da pochissimi tracker. Tuttavia però qualche tracker c'era e, presumibilmente a causa di uno di questi in comune tra i due siti, a un certo punto sono stati entrambi coinvolti nel cuore del grafico.
Ma anche in questo caso sto avanzando delle ipotesi, dal momento che è impossibile ricavare delle vere e proprie tesi analizzando solamente queste dinamiche

Una ricerca di Marjan Falahrastegar, condotta grazie alla collaborazione di altri colleghi della Queen Mary University di Londra, ha provato invece a risalire alla provenienza geografica dei tracker, sempre nella speranza di individuare qualche indizio che possa suggerire qualche elemento in più.
Tra i risultati raggiunti dal team di studiosi, alcuni sono in linea alle conclusioni a




2014
*La distribuzione dei tracker di terze parti nel mondo*

cui sono giunto usando Lightbeam, come per esempio rilevare che ci sono alcune influenti realtà online che hanno il 3rd party tracking come servizio aggiuntivo ai tanti che già offrono. Tuttavia, mentre la mia analisi aveva individuato principalmente Google, e in parte anche Facebook, Falahrastegar e soci hanno messo sul podio Google con oltre quaranta tracker esterni, Microsoft con 19 e eBay con 7. Considerando che la ricerca è di fine 2014 questi risultati non sono certo obsoleti, ma è possibile che la gerarchia non sia esattamente la stessa.
La parte più interessante dell'indagine è però l'analisi geografica. Utilizzando un servizio di proxy che fa credere ai server visitati di essere un computer locale, la ricerca ha evidenziato differenze sostanziali di presenza di tracker di nazione in nazione. Il più delle volte questo sembra



riconducibile alle leggi locali. Infatti in Australia, in Canada e nella maggior parte delle nazioni europee, in cui ci sono norme specifiche che regolano il tipo di informazioni che possono gestire i cookie e le modalità con cui l'utente deve essere avvisato del tracciamento, la presenza di queste entità è moderata e la loro distribuzione è omogenea.
Se in Europa spicca la Germania per il suo numero di tracker, decisamente più elevato rispetto alla media dell'UE, è perché, al momento della ricerca, i tedeschi non avevano ancora reso effettive le norme europee. Il caso è lo stesso per la Turchia, Israele e Cina, in cui sostanzialmente mancano delle leggi in merito o sono ancora inadeguate.
Anche negli Stati Uniti e in Russia il numero di componenti terzi è abbastanza elevato, anche se in questi casi più che mancare delle leggi, manca chiarezza e coerenza nel dettarle.
Completa il quadro il Medio Oriente il cui numero di tracker, di provenienza statunitense, pare sia più alto della norma; tuttavia l'argomento non viene approfondito, ma solo lasciato lì, aperto a speculazioni.

# big data

Partendo intuitivamente dal loro nome, i big data potrebbero suggerire a un grosso database. La definizione è di certo corretta, ma parziale, in quanto non rileva la complessità del concetto, né è tantomeno il suo aspetto più significativo.
Per Oracle, le caratteristiche principali di un big data sono quattro, denominate quattro "V": volume, velocità, varietà e valore.
Il volume è la quantità di dati: come insegna la statistica più voci si hanno, più il valore tenderà a riflettere quello reale.
Come velocità si intende il rateo di ingresso di nuove informazioni e la capacità di elaborarle subito; molti servizi devono fornire soluzioni in tempo reale, ma senza una gestione immediata dei contenuti questo è impossibile.
La varietà è una caratteristica tipica dei big data, che si trovano a gestire diversi tipi di contenuti per ogni voce; un problema frequente è quando una fonte di dati cambia improvvisamente e senza preavviso la forma dell'output, poiché questo viene immagazzinato e utilizzato senza successo finché non viene corretto l'algoritmo di interpretazione.



Infine c'è il valore, il motivo per cui si usano i big data e sembrano avere così tanto successo.
Le statistiche, se usate impropriamente, rischiano di essere ingannevoli o addirittura fuorvianti; tuttavia, se il campione utilizzato è abbastanza grande e le relazioni sono state studiate correttamente, esse sono in grado di andare vicino alla rappresentazione della realtà. Se tiriamo una moneta dieci volte, è tutt'altro che impossibile che esca sette volte croce, ma sappiamo bene che la probabilità che esca testa non è del 30%. Se tiriamo questa moneta diecimila volte, non è detto che il numero di volte in cui è uscito testa sia uguale a quello di croce, ma di sicuro sarà molto vicino al 50%, e cioè la reale probabilità di ciascuna opzione. E più questo campione sarà grande, più la percentuale di ognuna tenderà a 50.
Attraverso algoritmi estremamente complessi è possibile rappresentare fenomeni anche molto articolati, fino al punto che compagnie di ingegneria sono in grado di prevedere quando un componente di una loro struttura sta per rompersi tenendo sotto controllo i trend di dati emessi da questa. In maniera simile, una compagnia di advertising può essere in grado di capire quando il sentimento del pubblico verso uno o più prodotti stia cambiando, e verso che direzione.

Questo è il principio dei big data: avere banche dati enormi che permettano di

avere una visione numerica della realtà molto precisa. Una vera e propria rivoluzione rispetto al passato.
Certo, la condizione è che a maneggiare queste complessità ci siano dei professionisti, in grado non solo di interpretarle nel modo corretto e di tenere in conto tutte le variabili del caso, ma anche di individuare delle relazioni significative tra voci di solito non confrontate: i cosiddetti insight.
Si parla di insight quando si rileva una nuova relazione tra due o più dati, come può essere un legame di proporzionalità; oppure anche quando l'osservazione di uno o più trend suggerisce al data scientist una possibile causa o conseguenza del cambiamento, in modo tale che questi vada a verificarla e, nel caso, a prendere i provvedimenti necessari.

Come si fa tuttavia a non perdere la bussola all'interno di una banca dati sterminata? Quando mancano gli automatismi ed è richiesto l'intervento umano per scovare degli insight, una tecnica abbastanza comune è quella del cosiddetto "piccolo incrementale" presa in prestito dalla biologia, in particolare da una teoria del 2002 di Stuart Kauffman. Secondo il ricercatore, è possibile che i piccoli organismi si evolvano in base all'adiacent possible, ovverosia che mutino verso sistemi più complessi attraverso piccoli cambiamenti incrementali. Questi piccoli step evolutivi richiedono un bassissimo



consumo di energia, ma possono di fatto innescare una reazione a catena che si riflette in una evoluzione significativa in poco tempo.
Utilizzando il medesimo concetto, è possibile per un data scientist innescare la generazione di nuovi e complessi algoritmi interpretativi, partendo dall'analisi di pochi, se non un solo, KPI (Key Performance Indicator), un valore considerato tra i più significativi del database.
Quando si incontra un dato interessante (seppur non ancora un insight), come per esempio un improvviso punto di rallentamento di un processo di produzione, si può unire allo studio l'andamento di un nuovo KPI alla volta, cercando di scoprire se un altro valore presenta variazioni di trend che possano essere collegate al primo. Se questo secondo cambiamento fosse per esempio proporzionale al primo, o se avvenisse nello stesso momento, o se comunque dovesse avere un altro motivo di esserne rapportato, ecco che potremmo aver trovato un insight. In certi casi, dal cambiamento di una tendenza, si trova un altro insight, e poi un altro ancora, finché si scopre che c'è un legame di interdipendenza tra un vasto numero di valori prima considerati irrelati l'uno con l'altro. L'importante è partire da alcuni KPI selezionati, per poi incrementarne gradualmente il numero.

Non è comunque raro trovare in rete opinioni sfavorevoli all'uso massiccio dei

big data, che possono essere ricondotte a tre principali ragioni.
La prima riguarda la questione etica del deposito, e del successivo utilizzo, di una grande quantità si dati personali, sebbene questi siano usati una volta anonimizzati, aggregati in statistiche e tendenze e mai associati a una persona o uno specifico comportamento.
Una seconda critica abbastanza diffusa è che le statistiche non possono rimpiazzare tutti i metodi analitici tradizionali sviluppati nei secoli.
La terza è infine che la quantità non vale la qualità.
Lasciando l'approfondimento della prima e sacrosanta critica, che riguarda più un problema etico di chi fa ricerca attraverso i big data e non tecnico, ai prossimi capitoli, ci soffermiamo ora sulle altre due.
Ritengo quest'utime obiezioni piuttosto blande, per non dire anacronistiche, che però sono state più volte usate da ricercatori e studiosi che si sono scagliati con forza contro i big data.
Riguardo al non poter sostituire i metodi già esistenti posso essere senz'altro d'accordo, e posso capire chi suggerisca un approccio cauto prima di cancellare anni di progressi in nome di una tecnologia che, per quanto estremamente promettente, non è altrettanto rodata. Detto ciò, non mi sembra un buon motivo per sconsigliare l'uso dei big data; tuttalpiù si possono integrare tra loro le due modalità. Inoltre, riguardo l'affermazione stessa, se



da un lato non è detto che rimpiazzeranno i vecchi metodi, dall'altro non è detto nemmeno il contrario; il tempo ce lo dirà.
Chi invece vuole difendere un dataset più ridotto, ma di qualità, usa come cavallo di battaglia il caso di Google Flu Trends(GFT). GFT era un sistema progettato da Google che, sulla base delle ricerche effettuate negli Stati Uniti, sarebbe dovuto essere in grado di individuare i trend dell'influenza, anticipando gli enti medici governativi. Ciò sarebbe stato possibile grazie a un aggiornamento in tempo reale dei dati, operazione non possibile ai centri statunitensi che avevano bisogno di aspettare alcuni giorni per ricevere le visite dei pazienti, e alcuni giorni per verbalizzarle e comunicare i risultati.
In un primo momento GFT parve funzionare, riuscendo a prevedere con particolare precisione l'andamento dell'influenza negli USA; tuttavia, in particolare nel 2011, il sistema incominciò a sovrastimare i casi, arrivando in certe settimane a prevedere il doppio delle visite di quelle che poi in realtà ci furono.
Anche nel 2013 prese uno spettacolare abbaglio, sovrastimando i casi di influenza del 140%.
La ragione di questa inconsistenza poteva essere di volta in volta diversa: in certi casi gli allarmismi dei media, che spingevano persone sane a cercano i sintomi della malattia, in altri il cambiamento di interazione della piattaforma Google stessa, e in altri ancora alcune dinamiche

del tutto casuali.
Per esempio, oltre a cambiamenti interni dell'ingegneria della ricerca, Google ha introdotto degli add-on con informazioni di salute prima dei risultati di ricerca, e ha aggiunto i suggerimenti in tempo reale. Sebbene queste funzioni fossero ottime per Google, hanno fatto sì che le abitudini di ricerca degli utenti cambiassero, e con esse la precisione di un algoritmo progettato su interazioni false (o meglio, non più vere).
Per tornare alle ragioni della critica, è vero, quantità non significa qualità e le stime clamorosamente sbagliate di Google ne sono la prova. Tuttavia bisogna considerare che la tecnologia è ancora in evoluzione, e non è in uso da molti anni come altre più rodate, ma nonostante ciò si è rivelata estremamente precisa per periodi di tempo di circa due o tre anni, dopo i quali tendeva a sbagliare e a dover essere riprogrammata. Bisogna solo stare attenti a non peccare di "big data hybris", come la chiamano Lazer e Kennedy di Wired, vale a dire di credere ciecamente ai big data perdendo di vista la realtà; e bisogna ricordare che i big data sono un'espressione della realtà, non viceversa. Detto questo, l'uso dei big data si è rivelato tanto efficace e rivoluzionario in tanti ambienti diversi che sembra paradossale metterne in discussione la validità.

A ogni buon conto, è interessante verificare come l'esempio di Google Flu Trends



rispetti pienamente le quattro "V" di Oracle: il volume sono tutte le ricerche Google di 300 milioni di persone, poiché inizialmente il progetto è partito nei soli Stati Uniti; la velocità di ingresso era in tempo reale, poiché i grafici si aggiustavano in automatico a ogni query ricevuta da Google negli USA; il problema della varietà è stato quello che ha introdotto errori nelle stime, poiché la forma di output del box di ricerca era cambiata per via del cambiamento dell'input degli utenti nel box di ricerca. Infine il valore: attraverso lo studio dell'algoritmo giusto, GFT poteva offrire con straordinaria precisione un servizio che lo stesso Stato americano non era in grado di dare se non con giorni o settimane di ritardo.

# digital footprint e digital shadow

I concetti di digital footprint e digital shadow non sono tanto distanti tra loro, al punto che non è raro che generino un po' di confusione. Quello che hanno in comune è di essere il prodotto finale di tutti i processi esaminati fin qui, ovverosia i resoconti dell'insieme di attività di profilazione e aggregazione dei dati degli utenti.

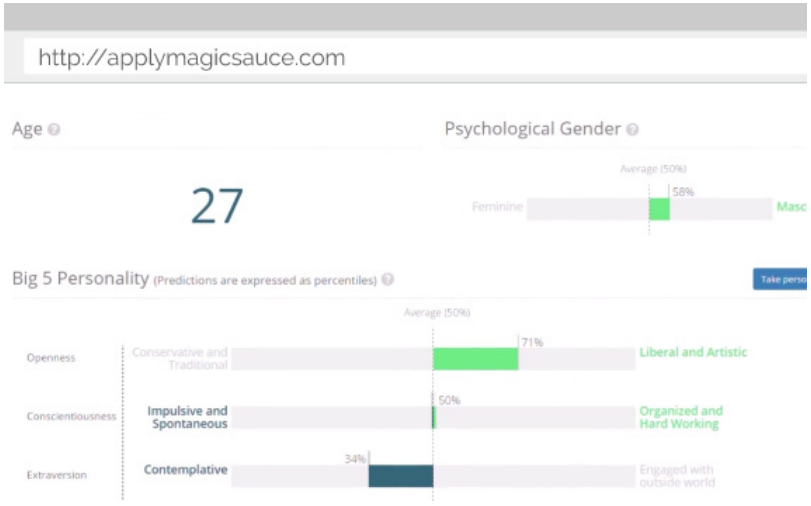
La digital footprint è sostanzialmente l'insieme di tracce che lasciamo dietro di noi in rete. Come abbiamo visto, parte di queste tracce possono essere lasciate involontariamente, come informazioni riguardo all'hardware che stiamo utilizzando, agli orari in cui siamo davanti al computer, ai luoghi che visitiamo se per caso ci dimentichiamo il GPS del telefono acceso. Altre volte invece, e questo è chiaro a tutti, lasciamo volontariamente nostre impronte: basta twittare quanto siamo annoiati di essere in università per dire dove siamo, mettere un like su Facebook per comunicare cosa ci interessa o condividere una foto con gli amici su Instagram per far sapere a tutti chi frequentiamo. Sono scelte volontarie e, salvo distrazioni, scegliamo noi il pubblico a cui mostrarle: solo agli amici per esempio, così chi non conosciamo non le vede. Altre volte ancora, lasciamo nostre tracce inavvertitamente. Questo significa che siamo ben consapevoli di pubblicare dei contenuti, ma non siamo in grado di valutare quanto pubblici diventeranno. È per esempio il caso di quando lasciamo un commento su una pagina pubblica: magari siamo consapevoli che diventerà pubblico, ma in genere non lo riteniamo un problema perché, a meno di ricevere centinaia di "Mi Piace" e finire nei commenti migliori, sappiamo che si perderà tra altre centinaia.

Esistono tuttavia degli strumenti di data crawling in grado di recuperare e raccogliere tutti i contenuti pubblici di un utente (o di una pagina), inclusi quindi i commenti; e questi strumenti non sono solo a disposizione del sito host, né soltanto ai programmatori che riescono a creare questo tipo di script (e di conseguenza, se lo pubblicano, anche a tutti gli utenti che lo trovano in rete), ma addirittura molti social network pubblicano per primi le API per farlo, come abbiamo visto parlando della user experience.

Alcuni tipi di dati che generalmente finiscono in una digital footprint possono essere i domini posseduti e gestiti,

i contenuti caricati su di esso (come testi, immagini e documenti), informazioni tecniche circa l'hardware e il software utilizzato per far stare in piedi quei siti, le informazioni dei social network quali età, sesso, provenienza, interessi e singoli contenuti (come pensieri, fotografie, opinioni e notizie), gli indirizzi email, i servizi ai quali quelle email sono registrate, le attività su quei servizi e naturalmente molto altro ancora.
Anche se la footprint è molto dettagliata e potenzialmente pericolosa, dal momento che potrebbe contenere a un certo punto qualche dettaglio come una password o un'informazione che doveva essere tenuta segreta, è anche vero che è molto difficile da navigare a causa dell'enorme quantità di elementi e la sintassi difficile da leggere senza una user interface progettata ad hoc. Gli stessi post scritti su Facebook rischiano di perdere di significato se presi singolarmente, lontano dai post immediatamente precedenti o successivi, oppure se privati dei commenti.

La digital shadow è la figura completa che viene fuori dall'analisi di questo insieme di dati, o di una parte di esso. Corrisponde all'idea che ci facciamo sulle altre persone: possiamo non conoscere esattamente i cantanti preferiti di ciascuna, o la specifica opinione su un tema di attualità, o i dettagli dei plugin con cui hanno costruito i rispettivi blog, ma basterà avere un'idea di che genere ascoltano, che lavoro fanno e se hanno un blog per avere un'immagine tutto sommato verosimile dei fatti. Se mediamente questo processo avviene conoscendosi, frequentandosi e discutendo, la digital shadow viene tracciata collezionando singoli elementi online; è un processo che in genere può fare un essere umano, e lo facciamo involontariamente quando capitiamo sul profilo social di qualcuno che non conosciamo e cerchiamo di immaginare che tipo sia, oppure anche un computer se programmato bene, come vedremo più avanti nel caso di Cambridge Analytica e Donald Trump.
Al contrario della footprint, probabilmente non contiene singoli dati particolarmente sensibili, ma al contrario è in grado di rappresentare un quadro completo e "umano" del suo proprietario. La digital shadow è un elemento molto rile-




*Applymagicsauce stima la digital shadow a partire dalla digital footprint del nostro profilo Facebook*

vante per gli hacker che usano il phishing. Ma anche questo tema sarà affrontato meglio nei prossimi capitoli.

# attori

In questo capitolo vedremo chi sono le compagnie online più attive e influenti nell'uso dei big data. Cercheremo di capire le differenze di trattamento dei dati da parte di compagnie simili, e di individuare chi si adopera per massimizzare la diffusione di dati e chi, al contrario, offre i propri strumenti per limitarla; se c'è qualcuno che aggira le regole e qualcuno che le fa rispettare.

# servizi online: il caso di Facebook e Google

L'uno è il sito con più utenti singoli iscritti, l'altro è per distacco il più visitato al mondo. Ma sembra che i due colossi americani, Facebook e Google, siano in testa anche per capacità di profilazione degli utenti, come sembrerebbero dimostrare diverse ricerche. Anche secondo Lightbeam, i due siti sono al centro della rete.

## FACEBOOK

In Facebook sono molto aperti sulla gestione dei dati: dichiarano a cuor leggero di utilizzare potenti strumenti di tracking. Negli ultimi anni hanno cercato di rendere sempre più chiaro il loro uso dei dati, attraverso il rilascio di numerosi aggiornamenti, cercando di ridurre sempre di più il "legalese", finché nel 2015 hanno lanciato il format "privacy basics". Si tratta fondamentalmente della pagina di data policy attuale, con informazioni molto chiare e semplici e una navigazione schematica, intuitiva e quasi piacevole.
Usano sempre il dato inserito e i suoi metadati associati, e sono in grado di raccoglierli da tantissime sorgenti diverse. Leggendo i loro termini del servizio, questo è ciò che comunicano di salvare:

**DATI DELL'UTENTE**

Dati immessi dall'utente: biografia, status, hashtag, località, persone taggate, etc.

Azioni effettuate: check-in, mi piace, etc.

Attività dell'account: orari, tempi di permanenza, post più seguiti, persone più seguite, persone più contattate, etc.

Dati di pagamento

Dati inseriti da altri: post in cui si è taggati, messaggi ricevuti, visite al profilo, etc.

## DATI DAL DISPOSITIVO

Dati sul device: fisso o mobile, modello, sistema operativo, componenti hardware, versione del software, etc.

Localizzazione: anche in background attraverso GPS, Bluetooth, WiFi, etc.

Connessione: operatore, lingua, indirizzo IP, numero di telefono, etc.

## DATI DA SITI ESTERNI

Siti con servizi e API Facebook: pulsante “mi piace”, “condividi”, sezione commenti, etc.

Siti partner: partner commerciali, advertiser, etc.

Siti proprietari: Instagram, Whatsapp, etc.

Sui metodi di raccolta dei dati sono molto chiari e onesti, non facendosi problemi anche a comunicare di tracciare in modo talvolta un po’ ossessivo dati che in certi casi possono essere di altissima sensibilità, come i dati della propria carta di credito o la localizzazione anche mentre il telefono è bloccato in tasca.
Dove invece potrebbero essere un po’ più



specifici, se non addirittura meno ermetici, è la sezione in cui spiegano cosa intendono fare esattamente con questi dati:

## USO DEI DATI

Fornire e migliorare servizi: suggerimenti di ricerca, tag, eventi interessanti, etc.

“Comunicare con te”: pubblicità, aggiornamenti dei Termini del servizio, informazioni su nuove funzioni, rispose a messaggi scritti all’assistenza

Monitoring dell’andamento di pubblicità, servizi, etc.

Sicurezza: rilevare violazioni dei termini, della legge, rilevare minacce, etc.

Mentre nella prima parte non risparmiavano esempi e dettagli sulle modalità di download dei dati, in questa parte mantengono un linguaggio generico; inoltre il secondo punto è come minimo poco chiaro. Passi la pubblicità, sulle cui modalità non scendono nello specifico, ma le altre ragioni per contattare l’utente paiono quasi una scusa, poiché basta un indirizzo email - peraltro necessario per iscriversi - per notificare un cambiamento sulla piattaforma o rispondere a una richiesta.

Il documento prosegue specificando con

chi e in che modo Facebook condivide i dati. Anche qui non si evita di entrare nel dettaglio finché si parla dell'uso interno, per diventare un po' meno precisi sul quando e come questi dati lascino i servizi Facebook.

### TRA I SERVIZI PROPRIETARI

Destinatari: audience del post, che potrebbe essere pubblica. I destinatari possono scaricare e condividere il contenuto, anche esternamente

Destinatari di chi ha pubblicato: audience di chi ha taggato l'utente

App, siti, integrazioni di Facebook o di chi ne usa i servizi: generalmente elementi pubblici, informazioni inserite nell'applicazione, altre informazioni previo il consenso

Prodotti proprietari

Eventualmente, al nuovo proprietario di Facebook

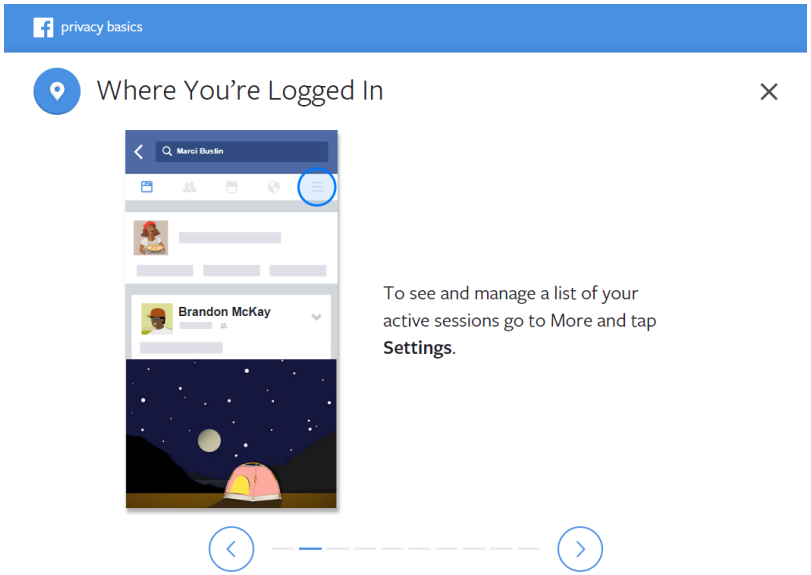
Per quanto riguarda partner e clienti esterni, Facebook condivide dati con:
. 3rd party tracker: inclusi quelli tecnici e di profilazione. Sono escluse le informazioni personalmente identificabili, tra cui età, località, etc.
. venditori, provider di servizi e altri partner: vengono definiti molto genericamente "partner esterni". Non è specificato se le informazioni siano personalmente identificabili oppure no. Sembra che possano dare qualunque informazione a chiunque tra i loro partner esterni, i quali sono però tenuti ad aderire a condizioni di trattamento strettamente confidenziali, comunque non specificate.



L'unico metodo suggerito per esercitare il controllo su tutte queste informazioni è usare l'activity log. Il documento non entra nel merito, tuttavia l'activity log permette di gestire solo le informazioni che si condividono a un pubblico. Per esempio non sembra possibile avere alcun controllo sulle informazioni che riguardano i dispositivi, i dati di pagamento e tutto ciò che non fa parte di quanto condividiamo con gli altri.
Facebook si riserva inoltre il diritto di mantenere le informazioni "finché ne abbiamo bisogno". Si impegna tuttavia a eliminare tutte le informazioni nel momento in cui uno dovesse cancellare il proprio account da Facebook.
In circostanze straordinarie si riserva il diritto di usare le nostre informazioni o, in caso di richiesta, di condividerle a terzi. Tra le possibili richieste sono inclusi procedimenti giudiziari, sospetti di attività illecite o frodi, situazioni di pericolo, sospetto di violazione di termini e condizioni. Infine, si riserva di mantenere attive le nostre informazioni per un minimo di un anno nel caso in cui il nostro account

venga sospeso per qualche violazione dei termini.

In ultima, Facebook fa parte di entrambi i "safe harbor" (USA-UE e USA-SVIZZERA). Dal momento che gli Stati Uniti, l'Unione Europea e la Svizzera hanno regolamentazioni diverse in fatto di privacy, gli USA hanno sviluppato, con l'UE prima, e con la Svizzera poi, due accordi separati che potessero soddisfare i criteri e le leggi sulla privacy di entrambi. Far parte dell'harbor significa che le condizioni di privacy e uso dei dati sono ritenute idonee a poter erogare il servizio in entrambi i paesi dell'accordo.




*Facebook Privacy Basics: un'interazione intuitiva per conoscere i termini del servizio*



## Cambiamenti nei termini del servizio

Le attuali "Privacy Basics" altro non sono che la vecchia data policy riscritta e riprogettata per essere più veloce e facile da capire per l'utente. È possibile che ci siano stati dei cambiamenti minori nel testo, ma in ogni caso non sono stati particolarmente rilevati, data la difficoltà nel reperire i dettagli.
Al contrario, si è scritto abbondantemente del cambio di forma e a proposito di altri cambiamenti passati, in cui non sono mancate polemiche.
Le privacy basics sono entrate in vigore nel gennaio 2015.

Nell'agosto 2013 un articolo su PC World mette in guardia su una imminente nuova privacy policy. Pare che anche qui si tratti principalmente di una modifica volta a riscrivere e a semplificare i vecchi termini e condizioni. In effetti vengono citati solamente due cambiamenti.
Il primo è che Facebook avrebbe aggiunto una funzione (a oggi non ancora disponibile a tutti) grazie alla quale il sistema sarebbe stato in grado di riconoscere gli utenti attraverso la foto del profilo, e quindi successivamente di individuarli in altre foto e suggerire il tag.
Il secondo cambiamento è che sarebbe stato esplicito che le informazioni passate agli advertiser fossero prive di tutti i dettagli personalmente identificativi.

Non è dato tuttavia sapere se anche prima la norma fosse questa, pur senza che fosse apertamente dichiarato.
Un articolo del Washington Post, contemporaneo al rilascio di questo aggiornamento, entra nel merito di una controversia non menzionata da PC World: Facebook aveva infatti aggiunto una riga in cui pareva dare per assodato che i minorenni godessero implicitamente del permesso dei genitori. Dopo varie proteste la riga è stata tuttavia rimossa.
Inoltre, in questa modifica erano presenti estensioni dei permessi sui dati di geolocalizzazione e di partnership con data broker.

Un post di Electronic Frontier Foundation (un'organizzazione no profit di attivisti per i diritti nel mondo digitale) discute un aggiornamento di dicembre 2009: in questa modifica venivano implementate numerose funzioni giudicate negative e un paio positive, tutte in vigore ancora oggi, oltre a qualcuna che nel frattempo è cambiata.
I giudizi positivi riguardano l'allora nuova possibilità di impostare il livello di privacy di ogni singolo post. Nonostante ciò EFF sostiene che si sia trattato di fatto di un provvedimento di facciata, con l'obiettivo di dimostrarsi aperti a migliorare le condizioni di privacy della piattaforma, ma che in realtà nel complesso andava nella direzione opposta.



**Facebook su Lightbeam**

Con una certa dose di stupore da parte mia, Facebook all'inizio occupa una posizione piuttosto marginale nel grafico, e l'unico dominio esterno di raccolta di dati a cui è associato è in realtà fbcn.com, a sua volta di proprietà del social network. Dopo le prime ore, a mano a mano che inizio ad aprire articoli su svariati siti, inizia un percorso che lo porta a trovarsi sempre più vicino al centro del reticolato globale generato da Lightbeam. Poiché i siti terzi di tracciamento - rappresentati come triangoli - sono una minoranza, sembra che Facebook non usi molti tracker esterni. Tuttavia i collegamenti con i cerchi - i siti visitati - sono sempre più numerosi, a indice del fatto che ci siano stati degli scambi di informazioni; questo può voler dire, o che Facebook ottiene dei report dai siti visitati, per esempio nel caso in cui questi siano partner oppure usino i social media button, oppure che siano questi siti a usare dei dati provenienti da Facebook per mostrarmi pubblicità mirate. È curioso notare come nessun sito o servizio di Google sia mai collegato a Facebook con l'unica eccezione di YouTube, nel quale di sicuro ci sono i social media button per le condivisioni.

Google, diversamente da Facebook, è molto meno chiara nel dare spiegazioni circa il proprio utilizzo dei dati. Innanzitutto utilizza molto più "legalese", al quale prova a rimediare mettendo a disposizione un glossario per la terminologia tecnica. Ad ogni modo, i dati che preleva sono i seguenti:

### DATI DALL'UTENTE

Tutte le informazioni comunicate in prima persona dall'utente: nome, età, dati di carta di credito, etc.

### DATI COLLEZIONATI DA GOOGLE

Informazioni sul dispositivo: componenti hardware, software, sistema operativo, etc. Se mobile, anche operatore, numero di telefono, etc.

Log di attività: query di ricerca, dati sulle chiamate, indirizzo IP, attività del dispositivo, cookie, etc.

Localizzazione: GPS, indirizzo IP, WiFi, etc.

Dati d'uso sui servizi: installazione, disinstallazione, check periodici, aggiornamenti automatici, etc.

Memoria locale: browser, cache, cookie, etc.



Come vediamo, Google è molto meno metodica e ordinata di Facebook nell'elencare i dati prelevati, mettendo in un'unica categoria dati come l'età e le informazioni di pagamento, oppure sistema operativo e numero di telefono.
Inoltre, anche leggendo gli esempi di dati prelevati, Google non sembra specificare esattamente tutto ciò che preleva esprimendosi con termini più generici. Anche quando ci si addentra in casi specifici si parla di funzionalità molto elementari, che certo non hanno bisogno di grandi quantità di dati per funzionare: per esempio, si dice che possono leggere dei dati tra cui i nomi dei contatti, così quando su Gmail si inizia a scrivere nel campo del destinatario un nome si possono vedere i suggerimenti.
Quando il discorso diventa un po' più complesso il linguaggio torna subito a essere tecnico, e benché ci sia sempre a disposizione il glossario, tante volte questo non risulta chiaro.
In una parte del contratto si dice che uno dei supporti dai quali traggono dati è la "archiviazione web tramite browser" senza spiegare che cosa sia e in che modo lo faccia; la voce sul glossario dice letteralmente che "è un modo che hanno i browser per memorizzare dati in locale. HTML5 aiuta". Oltre a non essere affatto esaustiva come spiegazione, sembra che tuttalpiù alluda ai cookie o alla cache, se non fosse che questi sono già menzionati separatamente. Un'ipotesi è che si tratti di

una funzione specifica di Chrome; ed è in effetti plausibile, considerando che in altre occasioni si fa chiaramente riferimento ad Android pur senza nominarlo - ma identificandolo invece come un generico "dispositivo mobile". Allo stesso modo, questo potrebbe essere un riferimento a Chrome nonostante la dicitura generica di "browser".

In relazione all'uso a cui sono destinati i dati, è prevista la massima integrazione tra servizi Google: il nome utente è uniformato (se hai Google+ apparirà sempre il tuo nome vero, anche su YouTube, Gmail, etc.), i dati dei profili sono visibili come risultati del motore di ricerca (compatibilmente con le preferenze del singolo) e i dati raccolti da uno dei servizi possono essere letti e sfruttati dagli altri.
Dichiarano di usare i dati per migliorare la propria offerta, che significa creare e modificare servizi, oppure renderli più personalizzati e fruibili. A tale scopo utilizzano i cookie e i pixel tag, che sono entrambi strumenti di page tagging.
Si riservano il diritto di usare i dati in paesi diversi da quello dell'utente, nonché quello di usare i dati in modi non previsti dal loro regolamento, previa richiesta di consenso.

Va detto che la data policy di Google diventa a tratti quasi indisponente: non è raro che ripetano concetti già detti pochi punti più in alto, o che facciano esempi



scontati, perdendosi in discorsi inconcludenti senza poi raggiungere il nodo delle questioni.

Riguardo alla memorizzazione della propria attività sull'account, Google lascia tuttavia grande libertà di scelta. Mette infatti a disposizione i seguenti strumenti:

**MEMORIZZAZIONE DI ATTIVITÀ**

Comandi attività: per scegliere che dati memorizzare e di che attività

Google dashboard: per modificare informazioni dell'account Google, comune a tutti i servizi

Preferenze sugli annunci

Privacy sui singoli contenuti. Se pubblica, questi diventano reperibili tra i risultati di ricerca

Possibilità di essere visualizzati negli annunci: per esempio, "Il tuo amico XY ha messo +1" nei news feed dei conoscenti sui post sponsorizzati

Google afferma successivamente che potrebbe «non rimuovere le informazioni dai nostri sistemi di backup». La motivazione con cui viene giustificata questa scelta è la volontà di proteggere l'utente in caso cancelli qualcosa per sbaglio.

Gli enti terzi che possono entrare in possesso di parte dei dati raccolti da Google sono:

**CONDIVISIONE A TERZI**

Chiunque, previa richiesta di consenso all'utente

Gli "amministratori di dominio": per esempio, chi configura un account Google Apps per conto di un team

Le aziende affiliate che trattano i dati per conto di Google

Chi di dovere per ragioni legali

Anche Google rispetta i Safe harbor USA-UE e USA-SVIZZERA

**Cambiamenti nei termini del servizio**

Nel settembre 2014 si parlava di pressioni dell'Europa a Google perché due anni prima quest'ultima aveva apportato dei cambi drastici alla data policy, ma senza rispettare le norme europee. La modifica aveva previsto l'unificazione di sessanta regolamenti distinti - ciascuno per un diverso servizio Google - e non era stata data la possibilità all'utente di non accettare.



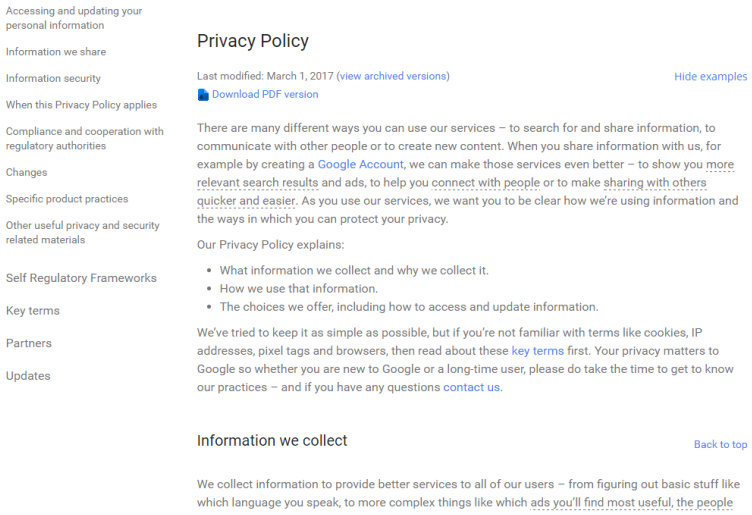
Anna Fielder di Privacy International la qualificò senza mezzi termini come una vergogna, criticando il fatto che a Google c'erano voluti ben due anni per presentare questa rivoluzione della policy che a conti fatti risultava inutile. Per Fielder, quelli che erano i nuovi termini del servizio non sarebbero dovuti essere altro che le sommarie linee guida della versione completa, delle fondamenta su cui semmai costruire; inoltre sottolineò che una casa come Google avrebbe dovuto applicare questi principi basilari ormai anni prima. E Rilevò infine che il linguaggio usato era talmente debole che l'impressione era che Google non avrebbe applicato niente di tutto ciò.
Questa modifica fu così controversa che pochi mesi più tardi, nel gennaio 2015, il commissioner del Regno Unito (al quale si unirono altri commissioner europei) ne discusse direttamente con Google, giungendo alla conclusione che fosse necessaria una revisione entro il 30 giugno con l'obbligo di specificare quali dati diventavano accessibili, con che modalità e da parte di chi.

**Google su Lightbeam**

Google si è posizionata subito come il nucleo del grafico, e lì è rimasta durante tutto il periodo d'uso. Dal cerchio col logo di Google (ho visitato www.google.com) si allacciavano la stragrande maggioranza dei siti che mano a mano visitavo, mentre

continuavano ad apparire dei triangolini (i tracker di terze parti) con a loro volta la favicon della G colorata, tutti dai nomi diversi. Spesso anche dei triangolini senza favicon si rivelavano in seguito essere di Google, traditi dall'URL. Google di per sé non è spesso presente sotto forma di triangolo, mentre è impressionante la quantità di quelli che possiede. Tanto per intenderci Facebook ne ha uno, come se un sito solo facesse tutto; quelli di Google vanno invece da Analytics a Google Fonts, fino a URL indecifrabili.

*La privacy policy di Google è meno user friendly rispetto a quella di Facebook, e anche meno chiara*






# browser e sicurezza

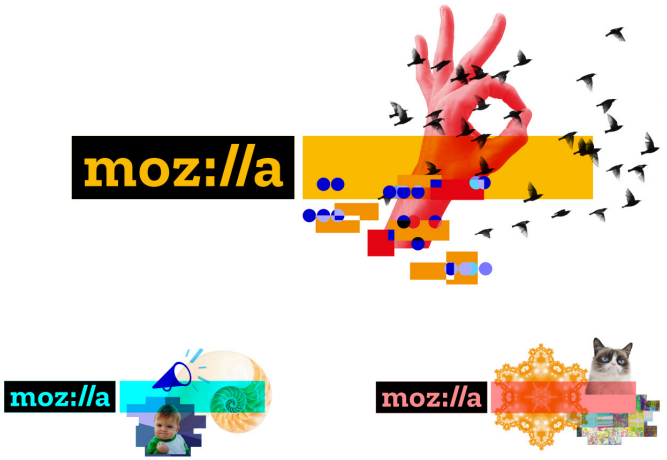
Prima dell'arrivo dei dispositivi mobili il browser era uno dei pochi punti di accesso a internet, e certamente quello largamente più utilizzato. Le specifiche funzioni di tracciamento, che sono per la maggior parte quelle descritte nell'unità 3, hanno prevalentemente bisogno di rispettare gli standard della rete, ragion per cui sono abbastanza simili in tutti i concorrenti.
Per questo motivo, oltre alla natura solitamente molto tecnica delle eventuali differenze tra le offerte, diventa interessante andare a vedere l'approccio delle singole software house per cercare di capire in quale direzione preferiscono dirigere i propri sforzi.
Non rientrano tra gli approfondimenti Internet Explorer, Edge e Safari poiché nessuno dei tre sembra intenzionato a offrire un prodotto con del differenziale - indipendentemente dal fatto che lo ricerchino nella gestione dei dati di navigazione o altrove - ma piuttosto ad accontentarsi della propria posizione di standard. A dimostrazione di ciò non solo

il fatto che siano preinstallati rispettivamente su Windows i primi due e su Mac OS il terzo, ma che una versione per il sistema operativo concorrente non esista nemmeno.

## FIREFOX

Tra i browser è certamente uno dei più attenti alla questione della privacy online. Mozilla, la software house che lo produce, è stata premiata come la compagnia online più affidabile del 2012 in fatto di privacy da Ponemon, un istituto indipendente di ricerca su privacy, protezione dei dati e sicurezza delle informazioni.
Oltre a tenere (sul blog e sul sito) alcune rubriche di informazione per l'utente molto improntate a renderlo consapevole dei rischi e delle contromisure al tracking dei dati di navigazione, ha anche sviluppato alcuni servizi e funzionalità interessanti legati a Firefox. Un esempio è sicuramente Lightbeam, di cui abbiamo già parlato a più riprese, e il browser è provvisto di impostazioni come la richiesta Do-not-track, implementata nel 2011 e da allora riprodotta da molti altri browser concorrenti. Si tratta fondamentalmente di una stringa di codice che, se attivata, ogni volta che si visita un sito manda una richiesta per chiamar fuori l'utente dall'essere tracciato nella sessione in corso. Purtroppo non può forzare la richiesta, e molti siti semplicemente rispondono negativamente, continuando






*Il logo di Mozilla è ancora più legato al web dopo il recente rebrand*

la collezione di informazioni; tuttavia, alcune associazioni di difesa dei diritti online si stanno battendo al fine di obbligare i servizi a rinunciare ai dati, qualora dovessero ricevere questa richiesta.
Inoltre Mozilla ha dato il via a Polaris, un'iniziativa con la missione di accelerare tutte quelle procedure e i progressi tecnologici finalizzati alla privacy online; aderiscono a Polaris anche altre entità importanti, dal browser TOR al CDT, entrambe decisamente coinvolte nella questione e di cui parleremo meglio più avanti.
La brand identity e il linguaggio di Firefox si identificano molto ai valori originali di Internet, tra cui openness, privacy e libertà: la pagina di download si intitola "Scarica Firefox - il browser libero e gra-

tuito"; nella scheda "about" della pagina Facebook mettono in chiaro la vision descrivendo internet come una risorsa che deve restare aperta e pubblica, e affermando che si adoperano a ciò lavorando con grande trasparenza.
Nella home page di Mozilla, dall'eloquente titolo "Internet for people, not profit", ben 2 link su 7 sono dedicati ai temi di sicurezza e privacy (uno porta alle iniziative di attivismo advocacy.mozilla.org sul criptaggio delle informazioni trasmesse, l'altro alla rubrica di insegnamento di cui parlavo prima). Più in giù il link a Firefox, "Impegnato a favore delle persone, della privacy e dell'open web".

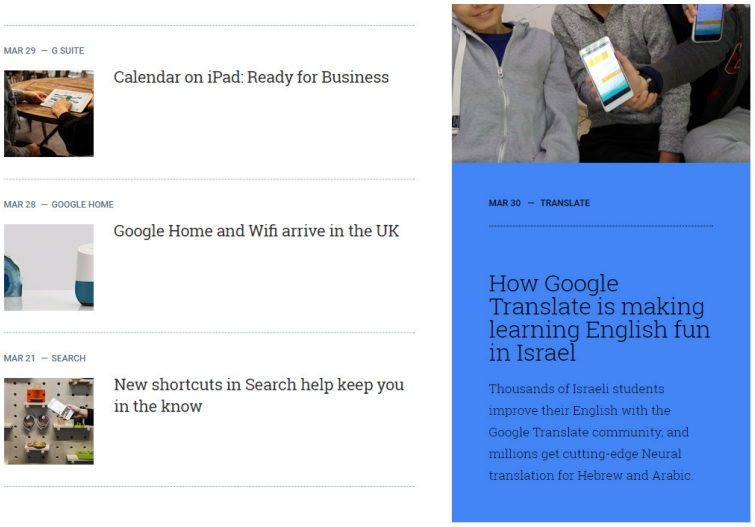
## CHROME

Essendo Chrome un prodotto di Google, non c'è una home page della casa di sviluppo sul modello di www.mozilla.org. Comunque, l'identità di Chrome è abbondantemente visibile già dalla pagina di download: la privacy e il tracking non sono menzionati, mentre si spinge molto sulla velocità e le prestazioni da un lato, e sulla semplicità e comodità di utilizzo dall'altro. Per esempio, ci sono delle preview di alcune funzioni come la barra degli indirizzi che è anche box di ricerca o il completamento automatico dei moduli. Viene presentato con la frase "Utilizza un browser web veloce e gratuito", diretto alle funzioni e alla semplicità. Nell'intera pagina di download, l'unica citazione a



privacy, anonimato ecc è in fondo al footer, sotto alla mappa del sito, insieme ad altre voci: "Google" - "Tutto su Google" - "Privacy di Chrome" - "Guida" - "Lingua del sito". La voce Privacy di Chrome porta a una pagina che assomiglia molto a un contratto, con centinaia di righe di testo nero su fondo bianco a caratteri piuttosto piccoli. Insomma, del tutto simile alla Data Policy di Google e ben poco user friendly.
Anche nei progetti correlati l'impostazione dei contenuti è simile: sempre nessuna menzione per i concetti di sicurezza se non quando si parla di protezione da virus, malware e altre minacce esterne; e in ogni caso l'accento è posto sempre sulle prestazioni. Succede così sulla home page di Google Code (piattaforma per svilup-

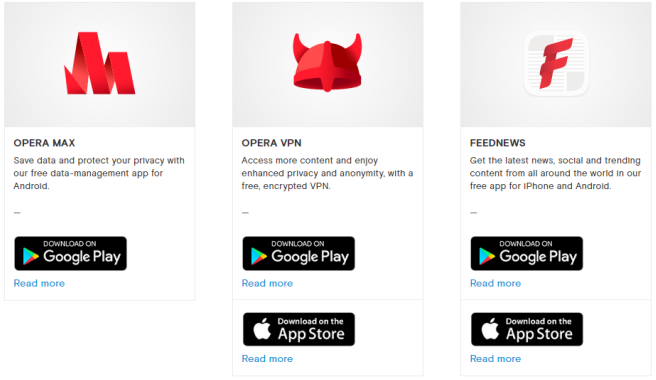
*Il blog di Google privilegia notizie sulla qualità e l'innovazione della propria offerta*



patori) e di Google Experiments; sul blog di Google le news riguardano sempre miglioramenti di prestazioni (in termini di velocità, compatibilità, batteria, protezione da virus...) e mai la privacy degli utenti.
Su Facebook, Chrome è semplicemente "a fast web browser built for today's web". Anche da una parola semplice come "today" si capisce la differenza col principale concorrente, Firefox. L'uno fatto per il progresso, l'innovazione e la prestazione, l'altro in nome dei valori originali e tradizionali del web.

## OPERA

La pagina di download cita tre caratteristiche principali, le solite tre comuni a tutti: "Un browser veloce, sicuro, facile da usare". Poi specifica "Prova il browser




*Due delle tre applicazioni del team di Opera hanno come obiettivo la privacy*



Opera - oggi con blocco della pubblicità integrato, risparmio energetico e VPN gratis"; ancora le magnifiche tre: un internet per tutti, prestazioni e privacy, questa volta sotto forma di specifiche funzionalità, con la prima che strizza l'occhio alla terza.
Su Facebook la sezione about non dice nulla, ma scorrendo i post in bacheca è evidente come ci sia attenzione verso il tema della privacy. Al momento, degli ultimi 20 post pubblicati addirittura 9 hanno temi come il cambio di IP, il VPN, l'ad blocking su mobile etc. Lo stesso avviene sul blog, in cui un significativo numero di articoli sono incentrati su sicurezza, anonimato e contromisure alla profilazione.
Comunque, il fatto di avere funzionalità come un ad blocker e un VPN integrati dimostra che siano molto consapevoli e impegnati sul tema e che siano interessati ad un'utenza altrettanto consapevole.
Significativo è anche il fatto che, oltre ai browser per computer e mobile, il team di Opera distribuisca anche due applicazioni: Opera Max e Opera VPN. La seconda assomiglia tutto sommato al VPN per browser, anche se qui la spiegazione del servizio è più sommaria e sembra servire più a sbloccare contenuti inaccessibili per ragioni di località e a valutare la sicurezza della linea, che effettivamente a depistare potenziali interessati alle attività. Opera Max, invece, è nata come app mobile per tenere semplicemente sotto controllo la

quantità di dati internet consumati, con la possibilità di comprimere il materiale scambiato (messaggi, video, immagini etc.) per ridurre i costi; tuttavia, successivamente la sua evoluzione è stata quella di tracciare quali applicazioni, in che momento e con che modalità trasferiscano dati verso terzi in background o senza un'azione precisa dell'utente, evidenziando così quali facciano tracking e in che misura. Di recente (a ottobre, 2016) sono state aggiunte alcune feature che bloccano certi componenti di terze parti (individuati grazie a un database comunemente usato dagli ad blocker) e che criptano i dati inviati quando si usano connessioni o protocolli non sicuri, come l'http.

### TOR

Mentre i prodotti menzionati fin qui, inclusi Internet Explorer, Edge e Safari, offrono generalmente una buona esperienza d'uso con l'aggiunta di funzionalità specifiche, Tor si propone innanzitutto come uno strumento di anonimato, cercando in seguito di rendere l'esperienza meno problematica possibile. Per questo motivo, non può essere considerato un vero e proprio concorrente degli altri browser.
Come si legge sulla homepage del progetto, Tor sfrutta una rete di utenti volontari. Quando ci si connette a un sito, anziché mandare le richieste direttamente



ai server, Tor le invia casualmente a uno di questi volontari, crittografandole. A sua volta, questo utente le invia a un secondo, e questo secondo a un terzo, e così via. A ogni passaggio è associata una diversa chiave di decrittazione, e alcune parti della richiesta restano distribuite nei nodi della rete. I dati assumono così una dimensione stratificata a cipolla (non a caso, TOR è l'acronimo di "The Onion Router") in cui solo il cuore possiede le informazioni, mentre gli altri strati, che sono i computer dei volontari attraverso cui è filtrata la richiesta, servono a crittare e a depistare eventuali tentativi di risalire ai dettagli della connessione.

## strumenti per gestire il tracking

Si può cercare di mantenere l'anonimato con diversi gradi di protezione. Tra i

sistemi più comuni c'è sicuramente l'uso di estensioni anti-pubblicità per browser, conosciute come ad blocker, e le più popolari sono supportate da almeno uno tra Chrome e Firefox.
Far uso di queste applicazioni non permette soltanto di evitare di essere disturbati da fastidiose animazioni, da banner che si sovrappongono all'articolo mentre si legge o dall'attesa di 30 secondi prima di vedere un video su YouTube; permette anche di andare a bloccare tutti quei canali che comunicano con l'esterno.
Se da un lato consideriamo affidabile il sito di Wired e lo visitiamo senza problemi, dall'altro non conosciamo il dominio dietro all'advertiser, il quale sfrutta un canale di apertura per inserire ciò che preferisce, tra cui un banner, un tool di 3rd party tracking, del malware. Ma ci si può anche spingere più in là con mail non tracciate, Tor e i servizi VPN.

C'è comunque una doverosa premessa da fare: sebbene alcuni metodi possano sembrare inattaccabili oggi, un giorno potrebbero essere bucati o raggirati. Inoltre, è bene tenere in considerazione che ogni catena è resistente solo quanto lo è la sua maglia più debole, e che a usare questi servizi, anche il più sicuro, ci sono sempre degli esseri umani. Se per una distrazione si lasciano tracce in grado di far associare un'informazione pubblica a un'attività anonima, si corre il rischio che qualcuno, unendo i puntini, possa risalire alla vera identità dell'autore di certe azioni, che siano le preferenze in fatto di abbigliamento o il consumo di materiale pornografico illegale, come dimostrato dal recente caso di arresto di un giovane newyorkese che attraverso Tor visitava siti di violenza sessuale e torture a bambini e animali.



Edward Snowden nel 2014 ha parlato al SXSW dando alcuni suggerimenti su come proteggersi dall'esposizione di informazioni su di sé, naturalmente con un occhio di riguardo nei confronti della sorveglianza della NSA. Per evitare di condividere i propri dati di navigazione è stato suggerito l'uso del browser Tor e di alcuni add-on, in particolare NoScript e Ghostery. Per nascondere i dati personali memorizzati sul computer ha suggerito invece di criptare l'intero disco fisso e la propria connessione con sistemi di "network encryption", come SSL.

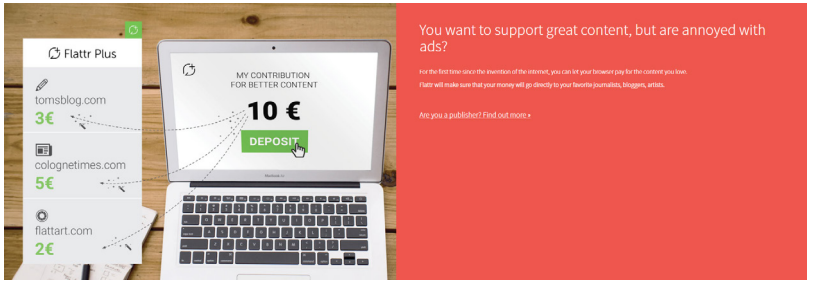
## GHOSTERY

Ghostery è un componente aggiuntivo per il browser che, come tanti altri, blocca le pubblicità. Dal momento che è stato consigliato da diverse persone influenti, tra cui appunto Edward Snowden, non metto in dubbio che funzioni benissimo.
Tuttavia non ho mai provato personalmente questo componente, a causa del suo modello di business, che è come minimo discutibile. Forse era diverso nel momen-

to in cui Snowden lo suggerì, o magari non era stato ancora reso noto.
Per guadagnare, infatti, Ghostery vende a grandi aziende (tra cui, si legge sul sito, Unilever, The Home Depot, Target, Toys R Us, etc.) tutti i dati d'uso del software, incluse informazioni sulle pubblicità bloccate, le modalità di blocco, le tecnologie usate e le pubblicità che invece hanno aggirato i controlli. Così facendo aiutano i grandi advertiser a fare ricerca su come eludere gli ad blocker in futuro.
Al di là del fatto che uno si schieri da una parte o dall'altra nel dibattito tra condivisione e privacy dei dati di navigazione, usare i dati con cui si proteggono i clienti per attaccarli più forte la prossima volta sembra come minimo un controsenso, se non un vero e proprio abuso.
Per il singolo utente è comunque possibile rifiutare di condividere i propri dati d'uso di Ghostery, che comunque verrebbero anonimizzati prima di essere condivisi con i partner. Nonostante questo, alcuni giornalisti hanno avanzato l'ipotesi che Ghostery venda i dati senza in realtà privarli delle caratteristiche identificative.

## FLATTR

Flattr è un servizio geniale nella sua semplicità, ed è un componente aggiuntivo per browser che serve ad aiutare i siti preferiti con delle donazioni. Basta associare un metodo di pagamento, per esempio PayPal, impostare la quota mensile totale che si è disposti a spendere, come 10 euro, e scegliere a che siti si vuole donare, diciamo Wikipedia, WikiLeaks e un blog di Tumblr. In questo modo, tutti i mesi Flattr preleva 10 euro dal conto PayPal associato e li distribuisce equamente ai miei tre siti preferiti.




*Flattr plus fa guadagnare i siti in base al numero di pagine visualizzate, ma senza infastidire con la pubblicità*

La comodità del servizio è che, una volta installato, si ha a disposizione nel browser un pulsante per aggiungere il sito corrente all'elenco delle donazioni. Inserito un quarto sito come destinatario, la quota mensile rimane di 10 euro, ma viene automaticamente ricalcolata perché sia assegnata a quattro siti anziché tre.
A maggio 2016 è stata anticipata l'uscita di Flattr Plus, che al momento è ancora in fase di sviluppo. Si tratta di un servizio che, fondamentalmente, permetterà un modello di business fai-da-te alternativo alla vendita di dati. Sempre allocando una quota mensile di donazioni, Flattr Plus blocca automaticamente le pubblicità come gli altri ad blocker e, in cambio, tiene un registro dei siti ai quali è stato negato l'introito e li risarcisce con una piccola quota del monte donazioni.

## ADBLOCK PLUS

Molto simile ad altri ad blocker per browser, e spesso confuso con il quasi omonimo AdBlock, è un comune add-on che blocca le pubblicità. Nella grande maggioranza dei casi ha successo, anche su siti importanti che presumibilmente possono contare su un buon livello di progettazione degli spazi pubblicitari. Nonostante la sua funzione primaria sia appunto quella di evitare le ad, ne offre anche altre interessanti. Permette per esempio di bloccare allo stesso modo i social media button, e cioè i pulsanti di condivisione e like, chiudendo così ogni possibile canale di passaggio di informazioni tra il sito visitato e un social network terzo. In questo caso è da segnalare che Adblock Plus evita efficacemente gli spot prima dei video su YouTube, mentre non è in grado di nascondere i pulsanti di condivisione esterna sotto al player. Offre inoltre la possibilità di inviare richieste do-not-track, di cui abbiamo già parlato nell'approfondimento su Firefox. Infine, una delle caratteristiche introdotte più di recente è la whitelist, che è un elenco di advertiser riconosciuti come sicuri e non fastidiosi dagli sviluppatori. Gli utenti possono quindi scegliere se bloccare tutte le pubblicità in assoluto, oppure se accettare di vedere almeno quelle proposte da questi pubblicitari selezionati. La whitelist è stata fortemente criticata da alcuni per diversi motivi, tra cui il fatto che gli advertiser considerati maggiori possono pagare per entrare a far parte dell'elenco. Se da un lato questo sembra scorretto, è vero anche che le compagnie più importanti non sono solite ricorrere a pop-up e animazioni moleste, ragion per cui chi scarica Adblock Plus con il solo scopo di non essere infastidito, non viene di fatto tradito dagli sviluppatori. D'altro canto si può capire il disappunto degli utenti che lo usano anche per ridurre il tracking da terzi dal momento che l'uso della whitelist è attivo di default. Ciononostante, è possibile disattivarlo in qualsiasi momento. Inoltre, degli introiti fissi garantiscono la possibilità che Adblock Plus resti costantemente aggiornato, che non significa solo un funzionamento più efficace, ma implica anche un guadagno in termini di sicurezza, poiché gli elementi più a rischio di malware sono i componenti software di un sistema. Se bucare un sistema operativo o un browser come Chrome può essere difficile, i componenti rimasti indietro con gli aggiornamenti come Flash e Adblock Plus diventano prede facili per gli hacker.

# organizzazioni per la privacy online

## PONEMON

Fondato da Larry Ponemon nel 2002, l'istituto fornisce assistenza privata ai business sui temi di privacy, sicurezza e in generale divulgazione dei dati, sia volontaria che involontaria (v. paragrafo sui data breach nel prossimo capitolo). Inoltre, pubblica annualmente delle ricerche, resoconti e classifiche considerati estremamente autorevoli e che godono di un buon livello di risonanza presso i media. Una delle classifiche più citate è quella sul grado di fiducia delle aziende in termini di privacy dal punto di vista dei propri utenti e/o consumatori, di cui esiste la versione integrale o suddivisa per categoria di appartenenza del business, per esempio banche, sanità, attività online.
Molto famoso è anche il report annuale sui data breach, che tra le tante informazioni include il costo medio di una perdita di informazioni, il costo medio di un furto di dati sensibili, la percentuale di rischio di subire un attacco o una perdita involontaria.



## CENTER FOR DEMOCRACY AND TECHNOLOGY (CDT)

Si tratta di una organizzazione non-profit statunitense (con sedi a Washington DC, San Francisco, Londra e Bruxelles) con il fine di trovare soluzioni tangibili alle principali sfide di internet alla privacy e alla sicurezza. La loro vision si fonda sui concetti di libertà civile e diritti umani. Si dichiarano totalmente imparziali, e i fondi provengono da fonti diversificate e a loro volta imparziali (la cena annuale di confronto, aziende finanziatrici di organizzazioni non profit, etc.).

## SAFE HARBOR

Già menzionato all'interno degli approfondimenti delle data policy di Google e Facebook, i Safe Harbor sono due accordi internazionali stipulati l'uno tra USA e Unione Europea, l'altro tra USA e Svizzera. Entrambi questi accordi contengono le norme che una compagnia online è obbligata a rispettare per poter essere considerata idonea a esercitare la propria attività in tutti i paesi coinvolti dal patto.

# rischi dell'uso improprio dei dati

Sebbene alcuni degli attori in gioco possano approfittare di situazioni nuove e poco chiare dei big data, è vero anche che, proprio perché parliamo di tecnologie e scenari senza precedenti, è ancora difficile mettere dei paletti chiari nella corsa ai dati personali; il discorso è applicabile sia in ambito morale che legale.

Certe volte il comportamento scorretto di alcuni è palese: è famoso il caso di Facebook uscito nel 2014, in cui fu rivelato che nel 2012 condusse segretamente un esperimento manipolando i news feed di 689 mila utenti per verificare se ci fosse la possibilità di influenzare l'umore delle persone mostrando una selezione di contenuti di felicità, tristezza o rabbia. Lo studio concluse che, sì, è possibile. Naturalmente questo ha attirato l'indignazione e lo sgomento di molti, tra i quali anche i redattori dell'analisi, dando orig-

ine a interrogativi circa la correttezza e la legalità di un'attività di questo genere. Per esempio, qualcuno si è chiesto cosa succederebbe se la CIA facesse pressioni a Facebook per aumentare il malcontento in Sudan con il fine di incitare alla rivoluzione. Qualcun altro si è chiesto se sia possibile sfruttare questa conoscenza per influenzare un'elezione: oggi sappiamo che è così, e nel prossimo paragrafo vedremo esattamente come.

Ci sono tuttavia diverse situazioni che rimangono dubbie, in cui prendere posizione non è così semplice; tante volte, nemmeno i ricercatori sanno bene come muoversi. Spesso infatti l'indignazione e la collera sono tra i motori più potenti dei processi che vanno via via a indirizzare sulla giusta strada l'etica di una disciplina. Ma se per branche come la medicina abbiamo alle spalle un'intera storia di proteste e scandali, tanto è vero che oggi nelle università esistono delle review board apposite, per una disciplina neonata come la big data analysis non ci sono ancora stati abbastanza bivi che indicassero un percorso chiaro.

È capitato che venisse chiesto alle IRB (Institutional Review Board, gabinetti universitari con il compito di valutare la correttezza etica di una ricerca) di giudicare delle ricerche sui big data; tuttavia, data la novità della materia, non esistono ancora comitati di esperti, di conseguenza le valutazioni vengono normalmente richieste a specialisti di altri settori. Per



capire com'è andata in questi casi, sono emblematici i risultati dei sondaggi presso i ricercatori di big data: prima di tutto, pare che tra loro stessi ci siano evidenti divergenze di opinione sulle questioni etiche. Inoltre, diversi ricercatori hanno riscontrato che la maggior parte dei dubbi fosse stata sollevata da loro pari - come amici nel settore e colleghi, oppure attraverso la peer review - e raramente dalle IRB.

Il designer che progetta un sistema di interazione che coinvolge la trasmissione, la collezione o l'elaborazione di dati deve necessariamente tenere conto dei rischi e, soprattutto, dei possibili usi del suo prodotto, anche quelli non previsti; con un approccio progettuale consapevole si possono evitare situazioni in grado di compromettere l'usabilità e la fiducia del supporto.

# usi moralmente controversi

Un argomento dei più controversi in tema di correttezza è l'influenzamento dei voti in politica. Sebbene fare campagna elettorale sia abbondantemente consentito, la possibilità di profilare e targetizzare i messaggi elettorali si trova in una zona grigia tra la pubblicità e la manipolazione, concettualmente difficile da inquadrare. La svolta in questa pratica è avvenuta negli scorsi mesi con due episodi chiave: la vittoria del leave EU nel Regno Unito e di Donald Trump negli Stati Uniti. Non è una novità che i social network e i big data vengano usati all'interno di campagne elettorali, poiché possono riprodurre le ricerche socio-demografiche che hanno sempre caratterizzato la politica moderna, anche prima del digitale, ma in modo molto più rapido, preciso e capillare. La novità è stata la modalità con cui sono state utilizzate le potenzialità dei mezzi.

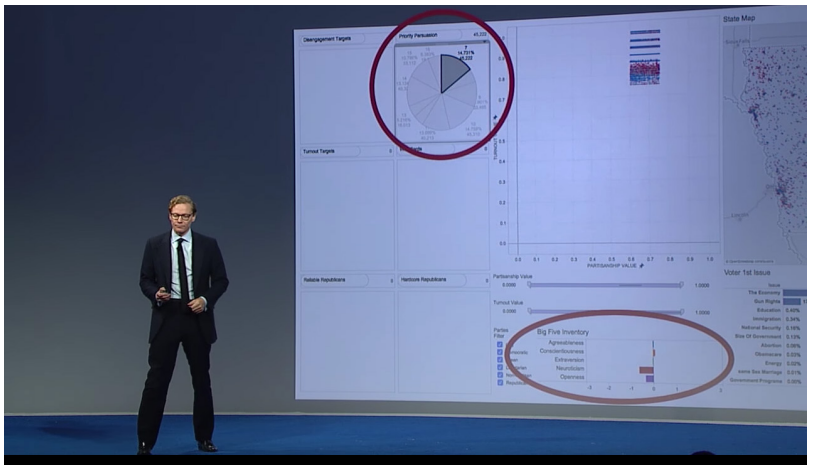
Nel 2008, gli allora studenti di psicologia a Cambridge Michal Kosinski e David Stillwell inventarono un'app per Facebook che riproduceva un quiz di personalità. I due ricercatori si aspettavano di ricevere risposta da qualche decina di amici e conoscenti, ma ben presto il quiz si diffuse e venne tentato da migliaia di persone, che in seguito diventarono milioni. La possibilità per gli utenti di acconsentire all'invio delle loro informazioni del profilo ai ricercatori fece sì che il duo si trovò per le mani un database smisurato, e decise di sfruttare l'anomalia della situazione a fini di ricerca; i dottorandi cominciarono infatti a prendere i risultati dei test di personalità e a compararli con alcune caratteristiche del profilo, come i "mi piace", il numero di contatti, il genere, l'età e i contenuti condivisi in bacheca. Ben presto emersero dei pattern: ad esempio, gli uomini con il like alla pagina dei cosmetici MAC avevano un'alta probabilità di essere omosessuali, mentre era un buon indice di eterosessualità quello al Wu Tang Clan; gli estimatori di Lady Gaga tendevano ad essere estroversi, della filosofia introversi; e così via. La ricerca andò avanti per anni, dimostrando che sulla base di soli 68 "mi piace" fosse possibile stimare il colore della pelle con il 95% di precisione, l'orientamento sessuale con l'88% e l'appartenenza al partito Repubblicano o Democratico con l'85%.

La ricerca fu pubblicata nel 2012 e subito notata; Kosinski venne immediatamente contattato due volte da Facebook,



una per un'offerta di lavoro, e l'altra per una denuncia. Dopo poco, anche un certo Aleksandr Kogan lo contattò, chiedendogli di acquistare l'accesso al database dello studio. Con una breve ricerca, emerse che Kogan lo contattava per conto di SCL, azienda proprietaria di diverse compagnie coinvolte nell'influenzamento di campagne elettorali nel mondo. Tra queste, alcune vantavano successi come l'aiuto al monarca in Nepal ai danni dei ribelli e influenze sulle elezioni in Ucraina e Nigeria. Più recentemente era nata una nuova compagnia in SCL, dal nome Cambridge Analytica.

Kosinski non voleva essere coinvolto in attività politiche di questo genere, pertanto rifiutò. Tuttavia, solo un paio d'anni più tardi fu costretto a sentire parlare ancora di Cambridge Analytica: questa infatti era stata la compagnia che aveva gestito la campagna elettorale per il leave che ha portato alla Brexit. Non solo: tra i loro clienti c'era anche Ted Cruz, un politico americano del partito Repubblicano che, come evidenziano dei dati divulgati da Cambridge Analytica stessa, riuscì ad ottenere dei risultati strabilianti sotto la loro conduzione: inizialmente, meno del 40% degli americani l'aveva sentito nominare, tuttavia in pochi mesi diventò il principale antagonista di Donald Trump all'interno del partito. Tuttavia, il suo ritiro dalla campagna elettorale nel maggio del 2016 fece sì che il suo rivale si trovasse da solo nella corsa alla candidatura, e che




2016
*Alexander Nix mostra la strategia data-driven usata per Trump e Leave.EU.*

*Nella pagina precedente, Michal Kosinski*

Cambridge Analytica passasse a lavorare per lui.

Conosciamo tutti il risultato del voto in America; le modalità con cui è avvenuta la digital strategy di Trump, invece, le ha spiegate candidamente Alexander Nix stesso, il CEO di Cambridge Analytica, presso il Concordia Summit, nel settembre 2016. Probabilmente ispirato dal lavoro di Kosinski e Stillwell, Nix dice che è ridicolo pensare di poter indirizzare un messaggio a dei gruppi demografici, come se tutte le donne bianche dai trenta ai quarant'anni dovessero votare per forza di cose lo stesso candidato. Loro infatti usarono dei gruppi basati sulla psicometria, proprio come il progetto del duo di ex studenti.
La compagnia acquistò e ottenne infatti dati da numerosissime fonti, complice anche la regolamentazione americana, che agevola questa pratica rispetto, ad esempio, all'Europa. Sulla base di questi, stando alle parole di Nix, fu possibile ottenere il profilo psicologico personalizzato di ogni singola persona adulta negli Stati Uniti, cioè approssimativamente di 220 milioni di persone.
Quello che hanno fatto a Cambridge Analytica, oltre all'incredibile lavoro di rifinitura dei dati, è stato preparare una digital strategy che andasse a colpire selezionati gruppi di persone. Nella maggioranza dei casi, queste erano raggiunte attraverso dei post su Facebook; stando sempre alle dichiarazioni di Nix, ne pubblicarono oltre 175.000 varianti. Dal momento che questi post erano mirati, risultavano invisibili a chi non rientrasse nel target preselezionato dalla compagnia; questa strategia è di fatto all'origine dell'incoerenza di Trump, che più volte ha espresso pensieri contrastanti con quanto dichiarato poco tempo prima, così come delle note affermazioni per nulla politically correct che lo hanno visto protagonista nei mesi di campagna elettorale. Nonostante qualche frase fosse riportata dai media tradizionali, generalmente il pubblico non previsto dalla progettazione di un post non poteva entrare in contatto con certi contenuti semplicemente perché non c'erano. Al contrario, affermazioni forti aumentavano il senso di appartenenza alla sua linea politica, sempre che ce ne fosse una univoca.
Il lato negativo di esprimersi in maniera

così cruda è tuttavia che si attirano antipatie, e si amplificano laddove già ci sono. Ma se c'erano dei gruppi di persone che non avrebbero mai votato per lui, allora sarebbe stato meglio evitare che queste persone andassero a votare del tutto. Con questo piano in mente, Nix e soci progettarono una serie di contenuti mirati che mettessero in cattiva luce anche l'altra candidata, Hillary Clinton. Per esempio venne mostrato a selezionati gruppi di afro americani un video - decontestualizzato - in cui la Clinton li definiva dei predatori.
Ma l'uso di questo big data non si è fermato online: in certi casi, gli agenti elettorali di Trump giravano di casa in casa per fare propaganda parlando con le persone. Grazie ai dati di Cambridge Analytica, potevano prevedere chi, all'interno di un condomio, potesse essere ricettivo nei loro confronti, e si preparavano una traccia della conversazione che più probabilmente potesse dare esito positivo prima di suonare alla porta.

Il caso è così complesso, e con implicazioni talmente forti, che ci si chiede dove sia la linea di confine tra ciò che è accettabile e ciò che non lo è. Quando arriveranno le IRB con competenze in fatto di big data, queste potranno giudicare le iniziative universitarie e di ricerca, ma non quelle commerciali e politiche. D'altra parte, la maggior parte delle azioni di Trump e Cambridge Analytica sono state corrette:



i dati sono stati ottenuti regolarmente, e la promozione sui social è permessa, comprese le funzioni di targetizzazione del pubblico. L'unica vera scorrettezza è stata quella di aver diffuso notizie false al fine di manipolare l'opinione pubblica; tuttavia anche questo è oggi un tema centrale di discussione, non ancora del tutto chiarito.

## data breach

Con l'espressione "data breach" si intende un evento in cui dei dati protetti, sensibili o confidenziali vengono esposti, copiati, trasmessi, visualizzati o maneggiati in qualsiasi modo da persone non autorizzate. Una fuoriuscita di informazioni riservate è considerata un data breach indipendentemente dall'origine dell'esposizione. Inoltre non sono rari i casi in cui questi siano stati provocati involontariamente; anzi, secondo Ponemon gli errori di natura umana sarebbero addirittura all'origine del 37% della totalità dei casi. Tuttavia, generalmente i casi più

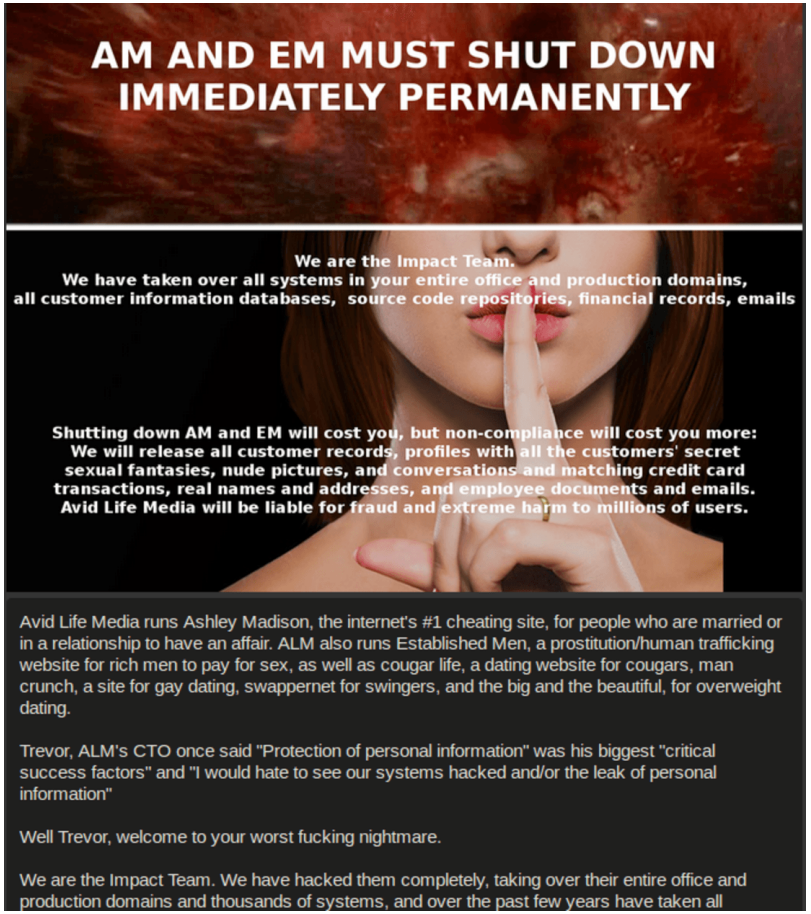
pericolosi e più gravi hanno origine dolosa, e alcune delle modalità attraverso cui avvengono possono includere il furto di dispositivi (quali computer e hard disk) e azioni mirate di hacking; inoltre si rivelano particolarmente rilevanti, anche mediaticamente, quando espongono contenuti personali o dati di pagamento. Tra i casi più noti c'è stato infatti il data breach di Yahoo!, avvenuto nel 2014 ma emerso solo due anni più tardi, in cui furono esposte informazioni personali come i dati di login; ad aggiungere risonanza certamente contribuirono la natura di Yahoo!, che sviluppandosi su molte piattaforme permette di conservare una grande varietà di informazioni dietro un'unica password, e l'entità dell'esposizione, che avrebbe coinvolto almeno 500 milioni di persone nel mondo. Altri casi molto noti sono quelli del PlayStation Network e di Target, rispettivamente nel 2011 e 2013, in cui furono rubati i dati di pagamento di 70 milioni di persone da ciascuno.

Un esempio che invece è giunto più difficilmente a orecchie europee, ma senza dubbio più spettacolare, è il caso del sito Ashley Madison. Sebbene oggi di definisca come un sito di incontri per appuntamenti discreti, una volta si dichiarava spudoratamente destinato ad incontri adulteri, al punto che il payoff del marchio recitava "Life is short, have an affair". Naturalmente questa natura suscitava non poche polemiche, anche se è



attivo ancora oggi dal 2001. Nel 2015, un gruppo di hacker chiamato Impact Team riuscì ad ottenere l'accesso a un database contenente i dati personali di 37 milioni di utenti; la portata dell'attacco fu così alta anche a causa della politica del sito, che preferiva mantenere in archivio i dati degli utenti registrati, al punto che la cancellazione dai server richiedesse il pagamento di una tassa da 15 dollari.
Sebbene l'attacco avesse esposto sia i dati di pagamento degli utenti, sia le credenziali di accesso, il vero obiettivo degli hacker era quello di ricattare i padroni della piattaforma chiedendo che venisse chiusa, poiché accusata di essere sbagliata e immorale. In caso di mancata soddisfazione della rischiesta, si dicevano pronti a pubblicare a poco a poco i nominativi di tutti gli iscritti, comprensivi di località e data di nascita, così che potessero essere scoperti dai rispettivi coniugi.
Poiché ALM, la compagnia proprietaria di Ashley Madison, rifiutò di cedere alle minacce, l'Impact Team cominciò effettivamente la pubblicazione di dati personali, che in certi casi includevano anche alcune chat private e lo storico dei pagamenti. In una seconda release vennero addirittura incluse delle email private del CEO stesso. Infine, oltre al danno si aggiunse la beffa: il data breach rivelò infatti che il sito non eliminava neanche i profili di chi aveva pagato per la cancellazione, esponendo così anche chi aveva smesso in precedenza di usarlo.

Naturalmente, le conseguenze furono pesanti. Ashley Madison e ALM ricevettero denunce e una class action in Canada, poiché affermavano esplicitamente di essere in grado di mantenere al sicuro i dati degli utenti. Inoltre il fenomeno fece spuntare dei ricattatori improvvisati, i quali sceglievano degli indirizzi email casuali dagli elenchi e chiedevano soldi in cambio di non mostrare i dati alle mogli, mariti e parenti. Una inchiesta di CybelAngel trovò 1200 utenti con un indirizzo email che terminava per .sa, indice del fatto che potessero essere provenienti dall'Arabia Saudita, dove l'adulterio è punito con la pena di morte. Ci furono anche dei casi di suicidio tra i membri nei giorni immediatamente successivi al leak, anche se uno fu valutato frutto di forte stress nella vita privata e lavorativa della vittima, e comunque nessuno fu considerato effettivamente imputabile alla vicenda.




*Il messaggio lasciato dall'Impact Team*

Anche se le gravissime conseguenze del data breach furono causate dalla natura del sito più che dall'attacco in sé, in realtà si possono correre pericoli analoghi anche conseguentemente ad eventi di portata minore. Non è raro infatti che degli hacker provino ad utilizzare le credenziali esposte da attacchi precedenti per provare ad accedere altrove, nella speranza che le vittime abbiano usato la stessa password. Inoltre, leak di servizi poco noti e non caratterizzati dalla sensibilità delle informazioni tendono a passare inosservati, per cui le persone esposte al rischio tendono a restarne inconsapevoli. Pertanto, per ridurre i pericoli sarebbe consigliabile evitare di usare sempre la stessa password, o perlomeno crearne una ad hoc almeno per i servizi che contengono i dati più sensibili, come estremi di pagamento o informazioni personali. O conversazioni adultere.

# sorveglianza e spionaggio di massa

Ma il rischio che i dati vengano usati impropriamente non è solo privato. È possibile che a compilare dei big data sugli "internauti" (termine che sta diventando desueto a fronte delle nuove modalità di accesso a internet, non più a sessioni limitate ma come collegamento costante) ci siano degli enti governativi, come hanno dimostrato, tra i tanti leak, le rivelazioni di Edward Snowden.
La sostanziale differenza con un tentativo di hacking privato risiede nei mezzi a disposizione degli attaccanti. Non solo possiamo presumere che un governo abbia a disposizione risorse ben più potenti della media, anche economicamente, ma soprattutto che può contare sulla propria posizione per richiedere, o addirittura pretendere, l'accesso al flusso di dati ai provider di servizi, per esempio le compagnie telefoniche e i servizi online.

Snowden è stato un funzionario della CIA e della NSA (National Security Agency, l'organizzazione per la sicurezza nazionale degli USA) per alcuni anni, durante i quali entrò gradualmente in contatto con alcuni dispositivi di sorveglianza dei servizi segreti. Egli fu colpito dalle potenzialità dei laboratori americani, ma non in positivo: secondo lui, buona parte della sorveglianza non era strettamente finalizzata alla sicurezza nazionale, e anche quando lo era veniva gestita con sufficienza, esponendo i cittadini americani a dei seri pericoli. Stando ad alcune sue interviste, l'episodio che sancì la sua definitiva risoluzione a rivelare tutto avvenne quando James Clapper, l'allora direttore dell'Intelligence americana, dichiarò sotto giuramento che "no, la NSA non raccoglie deliberatamente dati su milioni, o centinaia di milioni, di cittadini". Al di là della gravità del gesto, buona parte dello sconforto era dovuta al fatto che nessuno, nella NSA, fu sconvolto dalla menzogna. Citando uno studio di Hannah Arendt sui burocrati nazisti, inserì la reazione di semplice accettazione dei colleghi all'interno del concetto della "banalità del male": sarebbe a dire che, venendo in contatto quotidianamente con eventi di scorrettezza e malvagità, si riduce gradualmente la capacità di identificarli come tali.

L'oggetto principale delle rivelazioni di Snowden fu PRISM. Si tratta di un progetto segreto della NSA che, attraverso le richieste a compagnie online, tra cui principalmente Google, Microsoft e Ya-

hoo!, è in grado di far fluire attraverso i suoi sistemi tutto il traffico di dati interno e in entrata o uscita dagli Stati Uniti; secondo Snowden, una parte dei dati si limita a transitare dalle loro macchine, una parte viene brevemente conservata e un'altra ancora viene permanentemente salvata.

Le conseguenze di questo meccanismo, come emerge dai documenti esposti, hanno implicazioni morali e reali potenzialmente devastanti. Le prime vengono talvolta trascurate nel nome di una maggiore sicurezza, anche se è un'opinione diffusa che questa raccolta selvaggia sia in realtà inutile o addirittura controproducente. Dall'altra parte, però, Snowden stesso ha elencato alcune situazioni in cui questa pratica può rappresentare una grave minaccia. Per esempio, la NSA è solita passare un gran numero di conversazioni private all'intelligence di Israele senza preoccuparsi di anonimizzarle né di rimuovere i metadati; in queste conversazioni ci possono naturalmente essere dei messaggi di cittadini americani di origine araba o palestinese che scrivono ai propri parenti, rimasti in terre ora sotto il controllo degli israeliani; questo permetterebbe a Israele di individuarli.

O ancora, PRISM tiene d'occhio le abitudini sul consumo di pornografia da parte di attivisti e politici radicali, così da poter distruggere la loro reputazione anche senza vere accuse a carico. Questo caso in particolare ha ricordato a Snowden di



2014
*Edward Snowden*

quando, negli anni '60, l'FBI provò a ricattare Martin Luther King facendo leva sulla sua infedeltà e cercando di indurlo al suicidio.
I pericoli che corrono i cittadini sono quindi innumerevoli, e i documenti pubblicati sono solo una selezione di tutti quelli che sono stati consegnati ai giornalisti da Snowden.

È certamente giusto che un governo disponga dei mezzi per compiere un certo tipo di indagini, allorché fondate e approvate da un apparato giuridico; tuttavia non si pensi che non c'è modo di difendersi da un progetto governativo del genere: infatti la NSA, così come organizzazioni analoghe in altri stati, non compie queste ricerche in modo indipendente. Come si legge in un documento esposto dal Washington Post, circa il 98% della produzione di PRISM proviene da Google, Microsoft e Yahoo!, e anche il resto è originato da influenti compagnie che operano nel digitale e nelle comunicazioni, tra cui Facebook, Skype e Apple. Pertanto, nonostante avere il controllo sulla sorveglianza di massa sia difficile, è necessario conoscere lo scenario durante la progettazione di un artefatto, evitando che in questo siano presenti trasmissioni superflue di dati.



# social engineering 2.0

L'ingegneria sociale è una tecnica delle scienze sociali volta a influenzare il comportamento delle altre persone, ed esiste da prima del mondo digitale. Tuttavia il concetto ha assunto una nuova dimensione nel mondo della cyber security, dove si intende l'atto di indurre un essere umano a eseguire operazioni oppure a divulgare informazioni riservate con l'inganno. Il metodo più comune per raggirare le vittime è quello del phishing, ovverosia di inviare un link o un file contenenti malware nella speranza che vengano aperti. All'apertura, il malware si installa nel dispositivo attaccato e inizia a svolgere il proprio compito.
In tempi recenti la S.E. ha assunto caratteristiche nuove, tanto da poter parlare oggi di social engineering 2.0. Un po' come avvenne anche per il web, la "seconda versione" non è in realtà strettamente collegata con l'introduzione di una nuova tecnologia, o di una sua riprogettazione, bensì è dovuta a un radicale cambiamento delle tendenze d'uso, e di conseguenza anche delle sue potenzialità.

Una delle ragioni dell'esplosione recente della SE è l'alta solidità delle infrastrutture. Questa tendenza è certamente dovuta al grado di affidamento alla tecnologia del mondo odierno: tutte le compagnie hanno i propri archivi in digitale, si eseguono operazioni - come i pagamenti - online, e le persone conservano documenti e dati sui propri dispositivi; adesso inoltre sta avvenendo il passo successivo, che è l'uso del cloud storage.
Tra i tanti progressi in ambito tecnologico vanno considerati i sistemi operativi: mentre Mac OS negli ultimi anni non ha avuto grossi problemi, Windows in sole tre versioni ne ha risolti parecchi, basti pensare a tutti quelli per cui Vista, a suo tempo, fu sommerso di critiche; i dispositivi mobili invece hanno sempre avuto sistemi sicuri.
Un'altra componente molto migliorata è la navigazione da browser: oggi, grazie ad HTML5, è possibile programmare funzionalità, animazioni e interazioni che una volta erano realizzabili solo con Flash o altri componenti esterni. E questi componenti esterni costituivano - e costituiscono tutt'ora - un rischio, non solo perché si tratta comunque di un elemento perforabile in più nella catena dei sistemi di sicurezza, ma anche perché è possibile che non vengano aggiornati a lungo, così da diventare una preda più facile. Inoltre, oggi è diventato più comune che i siti internet usino una connessione con protocollo HTTPS, almeno in fase di



login. Si tratta di una connessione in cui i dati vengono criptati: pertanto, se anche qualcuno fosse in grado di intercettare il messaggio, questo risulterebbe illeggibile. Infine, abbiamo visto che ci sono dei componenti aggiuntivi, alcuni dei quali relativamente comuni al giorno d'oggi, che bloccando le pubblicità e i social media button chiudono la porta a canali aggiuntivi di trasmissione di dati.
C'è tuttavia stata un'effettiva innovazione tecnologica che ha quantomeno favorito il social engineering: oggi esistono dei sistemi di automazione che accelerano notevolmente il processo grazie a strumenti di object recognition, audio fingerprinting, OCR e altri modi di far capire alla macchina il contenuto semantico di un elemento testuale o multimediale.

Per tutte queste e altre ragioni, bucare un sistema di sicurezza è diventato impegnativo per un hacker. Per riprendere una metafora già usata, la sicurezza di un sistema è una catena che tiene chiusa la porta. Se una maglia di questa catena è fragile, non importa quanto siano resistenti le altre, basta rompere quella. Al giorno d'oggi, a causa di questo buon livello generale di sicurezza, la maglia più debole della catena spesso è l'essere umano.
Ma questa non è l'unica ragione dietro alla evoluzione della SE 2.0. Mentre una volta per raggirare l'essere umano si doveva tentare la sorte con generiche email di phishing, dalla percentuale di successo

bassissima, oggi grazie ai social network (e spesso a un uso poco responsabile di questi) capita che sia l'essere umano stesso a condividere informazioni che possono essere usate contro di lui. La disciplina di scegliere la vittima, studiarne la digital footprint e infine sferrare un attacco altamente personalizzato è detta "spear phishing". Solitamente questo avviene stabilendo un contatto privato, ad esempio via email, utilizzando un indirizzo o un profilo legittimo di un conoscente e facendo riferimento a fatti reali.
Un esempio verosimile potrebbe essere il seguente: un attaccante visita il sito web di una compagnia che intende attaccare. Cerca e scarica un file PDF che sia abbastanza recente, poi lo scorre cercando il nome di qualche collaboratore esterno; lo trova e lo conserva da utilizzare come falso mittente. Fatto ciò, individua una potenziale preda tra i dipendenti dell'ufficio nella pagina del team, possibilmente scegliendone uno che abbia il profilo LinkedIn. Sfrutta i dati, come la provenienza e il percorso di studi, per fare una ricerca filtrata su Facebook, e sul suo profilo scopre che due settimane fa era in vacanza in Toscana.
A questo punto, basterà creare un indirizzo email credibile, come ing.mario.bianchi@gmail.com e scrivere: "Buongiorno Carlo! Com'è andata in Toscana? Avete visto la Torre di Pisa? Chiedo scusa per il disturbo, ma mi sembra che nel PDF del vostro sito abbiate messo la vecchia versione del progetto. È così o sono io che mi confondo? Ecco il documento di cui parlo. A presto, Mario." L'allegato è un file .pdf apparentemente identico all'originale, ma contenente malware, e il gioco è fatto. Aggiungere gradi di dettaglio come il vero nome di un collaboratore, riferimenti a un lavoro effettivamente svolto insieme e addirittura a una vacanza possono far abbassare il livello di guardia su possibili incongruenze. Per esempio i due potrebbero non aver parlato della vacanza, ma diventa facile che la preda ipotizzi di essersi dimenticata di averne parlato, oppure che immagini che qualche altro collega l'abbia menzionata parlando del più e del meno. Inoltre, evitare di indirizzare il discorso direttamente all'interlocutore, dovendosi sbilanciare dando del "tu" o del "lei", aiuta a evitare di farsi scoprire.

Un attacco studiato in questo modo risulta così credibile che si stima che, mediamente, ci si accorga di essere stati colpiti a un anno di distanza dal fatto. Un tempo più che sufficiente per l'hacker di guadagnarsi l'accesso a più macchine e ai file o agli account desiderati. Inoltre, le percentuali di successo sono elevatissime: si stima che una persona su tre cada nella trappola cliccando su un link maligno, e che addirittura una su cinque sia soggetta a cliccare il link e poi anche a inserire delle credenziali nel sito fasullo sul quale viene condotta.

Sebbene il social engineering 2.0 non sia ancora famoso come fenomeno mainstream, la sua ascesa di popolarità tra gli addetti ai lavori è stata tale che oggi capita di sentirne parlare sia in fatti di attualità, che addirittura nella cultura popolare. All'inizio di quest'anno (2017) è salita alla ribalta delle cronache internazionali l'operazione di hacking "Eye-Pyramid". Due fratelli italiani, un uomo e una donna, hanno infatti ottenuto l'accesso ai computer di numerose persone e istituzioni governative ed economiche, tra cui anche personalità del livello di Matteo Renzi, Mario Draghi e Mario Monti, grazie ad un attacco di spear phishing. Nonostante il fratello sia riuscito a distruggere parte dei dati, è emerso che i due avessero l'accesso a documenti riservati di enti come la Camera dei Deputati e il Senato, e anche alle email personali delle persone attaccate.
Il metodo di attacco è stato proprio lo spear phishing: nell'episodio che li ha inchiodati, i due fratelli avevano scritto all'Enav identificandosi come uno studio legale in collaborazione col responsabile. Il funzionario che ha ricevuto l'email avrebbe dovuto girarla dunque al suo capo, ma poiché non gli risultava che ci fossero rapporti con uno studio con quel nome cercò di andare oltre, facendo ispezionare la posta al CNAIPIC (Centro



Nazionale Anticrimine Informatico per la Protezione delle Infrastrutture Critiche). Una volta scoperto il malware contenuto è stato possibile risalire ai server, ai quali erano associate le configurazioni di tutti gli altri PC collegati poiché colpiti precedentemente. I due sono stati alla fine arrestati; ciononostante, resta il fatto che con questa tecnica alcune delle personalità più influenti del mondo - parliamo dell'allora Primo Ministro della quinta economia europea e del presidente della BCE - sono state spiate per anni da due cittadini "comuni".

Ma gli episodi di cronaca non sono gli unici a far conoscere al pubblico le potenzialità del social engineering. Nella serie TV di successo Mr. Robot, che ha come protagonista un giovane e abilissimo hacker, sono diversi i momenti in cui siamo resi partecipi dei metodi che usa, che

2017
*Il sopralluogo della Polizia Italiana presso gli archivi dei fratelli Occhionero*

spesso non sono altro che dei giochi di astuzia, piuttosto che complesse operazioni informatiche.
Solitamente, nel cinema come in televisione, gli hacker non vengono mostrati in modo realistico, e la natura della loro attività non viene approfondita. Generalmente la capacità di hacking è rappresentata alla stregua di un superpotere, che garantisce incondizionatamente l'accesso ai sistemi oppure non funziona, a seconda di ciò che è funzionale all'evolversi della situazione. In Mr. Robot diventa essa stessa un modo di far evolvere la situazione.
Fin dalle prime puntate Elliot, il protagonista, usa tecniche di social engineering per carpire informazioni sui suoi conoscenti, il più delle volte per curiosità o comunque per avere gli elementi giusti per poter giudicare una persona o un evento. Ad esempio, hackerando la propria psicanalista scopre che questa è single e frequenta un sito di incontri, sul quale conosce un uomo. Questo non è chi dice di essere, trattandosi di un uomo sposato e dal nome diverso da quello con cui si fa chiamare sul sito. Elliot decide dunque di fare giustizia, e lo fa interamente con metodi di S.E. 2.0: prima segue la sua psicanalista mentre esce a cena con l'uomo. Dopodiché, quando questo prende da solo il taxi per tornare a casa, Elliot registra velocemente il numero di telefono della compagnia e la targa del mezzo, e telefona comunicando al centralino di aver



*Mentre Elliot (in primo piano) finge di chiamare la madre, il suo telefono si illumina in tasca*

dimenticato a bordo del taxi le chiavi di casa. La centralinista gli dice che può recuperarle direttamente presso il tassista, rivelando l'indirizzo di destinazione. Avendo così scoperto l'indirizzo della sua preda, vi si reca in un secondo momento aspettando che esca di casa. Ad un certo punto esce con il cane al guinzaglio; Elliot gli si avvicina, e chiede cortesemente di poter utilizzare il suo telefono per chiamare la madre. In realtà chiama sé stesso, avendo poi cura di cancellarsi dal registro telefonico del cellulare ricevuto in prestito. Avendo così ottenuto anche il numero, qualche giorno dopo lo chiama spacciandosi per un operatore della sua banca, scoperta sempre con tecniche di hacking; gli fa alcune domande con la scusa che siano per fini di sicurezza, e spera attraverso queste di venire a conoscenza della sua password oppure di poterle usare come risposta alla "domanda segreta" per

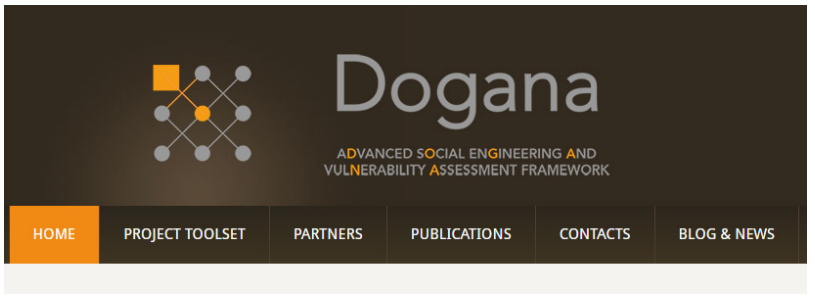
poter accedere senza credenziali. Evidentemente ci riesce, poiché in seguito torna sotto casa dell'uomo e lo minaccia di svelare alla moglie dei segreti compromettenti se non avesse lasciato la sua psicanalista.

# dogana: all'interno di una digital footprint

Dogana è un software di sicurezza informatica che permetterà di ottenere e lavorare digital footprint di aziende e persone; l'obiettivo è di eseguire perizie approfondite circa i rischi derivanti dai dati reperibili in rete. Attraverso lo studio e la progettazione del software, si fa chiarezza sugli elementi di cui è composta una digital footprint, quali sono gli strumenti di automatizzazione che aiutano gli hacker e di conseguenza forniscono spunti per una gestione più responsabile di ciò che si lascia sul web.

Il progetto Dogana nasce dall'esigenza di prendere contromisure contro il dilagante fenomeno del social engineering 2.0 mirato alle aziende. Nonostante esistano e siano conosciuti diversi programmi che possono in parte aiutare a difendersi dal SE 2.0, si tratta spesso di strumenti parziali, e quindi eccellenti in un compito, ma inaffidabili su tutti gli altri, oppure con grossi limiti sulla quantità di dati che riescono a gestire. L'obiettivo a cui mira il progetto Dogana è dunque quello di diventare una struttura definitiva in grado di affrontare il problema da cima a fondo, andandosi così a posizionare come leader del settore.
Per riuscirci è stato deciso di prendere spunto dall'offerta già presente online che, nonostante la sua frammentarietà, è capace di fornire a tale scopo esempi molto validi.
L'aspirazione sarebbe di assemblare esclusivamente componenti open source pre-esistenti all'interno di un'unica UI; nella probabile eventualità in cui alcuni elementi non fossero disponibili in open source, o che comunque non siano in grado di dare sufficienti garanzie, c'è anche






*La home page del Dogana Project (www.dogana-project.eu)*

l'intenzione di chiedere, ed eventualmente acquistare, i permessi per utilizzare motori e codici di software a pagamento.
Un team di programmatori che collabora al progetto ha infine il compito di unire tutti questi elementi e farli comunicare. Nel caso in cui qualche parte del sistema non fosse reperibile in alcun modo, si provvederà a progettarla ex novo.

Dogana è ancora il nome in codice per gli addetti ai lavori, pertanto il software risultante potrebbe avere in futuro un'altra denominazione; ad ogni modo, Dogana è una creativa selezione di lettere prese dalla sua definizione: advanced social engineering and vulnerability assessment framework.
Come da descrizione, il software sarà una struttura autosufficiente in grado di produrre perizie e attestati di rischio per i business. Per autosufficiente si intende che probabilmente non ci sarà uno staff di Dogana che farà funzionare il software ed eseguirà perizie su commissione, ma il programma verrà venduto o noleggia-

to. Ancora non ci sono dettagli precisi su quello che sarà il modello di business, se quindi fornirlo alle aziende direttamente oppure alle compagnie di IT che potranno utilizzarlo come strumento per eseguire a loro volta perizie ai clienti.
Quindi, nonostante il target a cui Dogana fa riferimento siano utenti esperti (anche nel caso in cui lo acquisti un'azienda ci si aspetta che a farlo funzionare sia il dipartimento di sicurezza), la progettazione deve tener conto che potrebbero non essere familiari con la navigazione di una digital shadow o con la data analysis.

Sebbene sia previsto che tra le funzioni rientrino anche attività di awareness, l'aspetto principale di Dogana sarà quello di poter accedere alla digital footprint e la digital shadow di una compagnia e di valutarne lo stato di rischio tenendo conto non solo dei dati pubblici dell'azienda, ma anche di quelli dei suoi dipendenti, poiché non è raro che le compagnie vengano attaccate proprio sfruttando un'ingenuità da parte di qualche impiegato. In seguito a questo tipo di valutazione, il software sarà progettato per poter preparare la simulazione di un vero attacco di spear phishing, di eseguirlo, e infine di scrivere un report sul grado di riuscita delle operazioni e sulle falle rilevate dal sistema.

Entrando invece nei dettagli circa il cuore di Dogana, il processo di esposizione dei dati e di simulazione di attacco avviene in



quattro fasi ben precise, ciascuna con un compito diverso e con un team di sviluppatori e professionisti dedicato. I nomi delle quattro fasi saranno certamente diversi nel prodotto finale.
La fase I per il momento si chiama IGAS, acronimo di Information Gathering and Analysis Service, e il suo obiettivo è quello di cercare, scaricare e riordinare i dati su una compagnia. L'elenco delle fonti da cui prelevare i dati non è ancora stata formalizzata poiché prima di tutto si vuole capire esattamente che tipo di dati si riescano a prendere, quali valga la pena utilizzare e quali no; terminata questa fase di scelta bisognerà capire, policy alla mano, quali servizi consentono il data crawling e con che modalità. Per esempio fare crawling su Facebook è molto semplice e in rete si trovano numerosissimi strumenti in grado di farlo, ma per i Termini del servizio si tratta di un'operazione illegale, anche se si ricercano solo file con livello di privacy pubblico. Tuttavia, vista l'importanza di Facebook nell'ipotesi di effettuare del social engineering sui singoli dipendenti, è in programma una richiesta per ottenere un'eccezione dall'azienda americana.
Al netto di tutto ciò, ci si aspetta che i dati prelevati saranno all'incirca i seguenti: per l'azienda si farà una ricerca completa sul dominio, su pastebin, sul DNS, su WikiLeaks ed eventualmente sui social; per i dipendenti, i social saranno invece una delle fonti primarie, ai quali segui-

ranno il dominio aziendale e le email.
Il passo successivo è di escludere tutti quegli elementi che non rispettano i termini e servizi di Dogana e della compagnia analizzata. Sebbene, a rigor di logica, per simulare una reale situazione di pericolo sarebbe bene conservare tutto e a maggior ragione sfruttare gli elementi dalla sensibilità più elevata, bisogna rammentare che essendo destinato a diventare un software istituzionale - come vedremo nel paragrafo successivo, è patrocinato dall'Unione Europea - deve rispettare regole piuttosto restrittive sulla privacy e la legalità, anche in considerazione del fatto che possa essere utilizzato come minimo in tutto il suolo europeo, e per questo deve attenersi a un discreto numero di normative diverse.
Successivamente, a ogni elemento andranno aggiunti dei tag per poterne rappresentare quantitativamente delle caratteristiche qualitative. Nonostante si sia ancora piuttosto lontani dal concretizzarne le modalità, è in previsione l'uso di alcuni strumenti di deep learning per poter partire da una base di tag automaticamente applicati dal software.
Infine, dev'essere possibile la navigazione dei dati risultanti, che saranno a questo punto una versione raffinata della digital footprint. Nelle fasi successive, e in particolare durante la preparazione dell'esca, si cercherà di tramutare selezionate parti di questa footprint in una digital shadow da poter attaccare.



La fase 2 è proprio quella di preparazione dell'attacco; la sigla, TAHP, significa infatti Tools for the Attack and Hook Preparation.
Accedendo alla visualizzazione della footprint, gli attaccanti (spesso chiamati penetration tester o, in forma abbreviata, pentester) navigano tra i dati, cercando un insight o un argomento da poter sfruttare per costruire un'esca efficace. Una volta individuata, sfrutteranno le email carpite da IGAS, oppure fornite dall'azienda per questo scopo, che le sceglierà in base all'obiettivo e che potrebbe essere un gruppo specifico di persone ritenute più vulnerabili, oppure un dipartimento che abbia un buon grado di prestigio nella compagnia da una parte e probabilità di essere ingannato dall'altra. Nonostante l'email sia uno dei mezzi più utilizzati nello spear phishing, è tutt'altro che l'unica via percorribile. Tuttavia, per ragioni di semplicità, al momento è l'unica contemplata da Dogana; sarà comunque possibile aggiungere altre modalità di contatto più avanti.
Attraverso la UI costruiranno quindi un'email assemblando dati che possano aumentare la verosimiglianza del testo, dai fondamentali come il nome e il genere della preda, fino a informazioni personali che sono emerse dall'information gathering, quali nomi di parenti, banche, compagnie assicurative, dettagli sugli autoveicoli, etc.

La fase successiva è ovviamente l'esecuzione dell'attacco, chiamata TEAT: Tools for the Execution of the ATtack. Si inizia con il classico pulsante rosso dei cartoni animati - "esegui attacco" - e si procede con l'analisi in tempo reale dell'andamento dell'attacco. La dashboard deve fornire tutti gli strumenti per valutare immediatamente l'efficacia, poiché in media ci si aspettano i primi risultati già nei primi minuti, se non addirittura in pochi secondi. Nel caso in cui qualcosa vada storto - per esempio nessun click nei primi cinque minuti - si può procedere con alcune operazioni tra cui abortire l'attacco, modificare i destinatari, etc. In un caso del genere è possibile che l'attacco sia stato individuato dai filtri anti spam; bisogna dunque interrogarsi sulle possibili motivazioni e tornare alla fase precedente per riformulare l'attacco e riprovare.

La quarta e ultima fase può entrare in gioco al termine dell'attacco oppure anche al termine della fase IGAS. Si tratta di TIAR, cioè Tools for Information Aggregation and Reporting. Il suo compito è quello di compilare i report che saranno consegnati al CEO della compagnia analizzata, o a chi di dovere, per mostrare il fattore di rischio e le falle dell'azienda. Naturalmente in questa fase è vietato mostrare i contenuti emersi durante la fase di information gathering per rispettare la privacy dei lavoratori, come non è permesso



svelare dettagli circa la costruzione delle email; né si intende in alcun modo colpevolizzare i dipendenti che hanno cliccato sul link, o peggio, inserito le credenziali nel sito trovato a quel link qualora TAHP ne abbia creato uno. Al contrario, TIAR produrrà dati quantitativi e grafici che mostrino i comportamenti positivi e quelli pericolosi, i risultati dell'attacco, la velocità (o la lentezza) con cui questo ha avuto successo, il numero di buchi aperti etc.
Poiché la simulazione d'attacco non è obbligatoria, il report potrebbe riguardare solo la fase di information gathering: in questo caso si evidenzieranno dati come la percentuale dei dipendenti a rischio, il numero di informazioni sensibili presenti in rete, la probabilità di subire attacchi a seconda delle fonti degli elementi ritenuti più pericolosi, etc.

## realtà coinvolte

Come accennato in breve nel precedente paragrafo, Dogana ha ricevuto il patroci-

nio dell'Unione Europea. Nello specifico ha ricevuto fondi per oltre 4,5 milioni di Euro nell'ambito del programma Horizon 2020, un'iniziativa per investire fondi in progetti di innovazione. Data l'entità del prodotto, sono numerose le realtà coinvolte nel progetto, ciascuna con un compito diverso, incluso quello di vittima dell'attacco.

Io ho fatto parte per alcuni mesi di una di queste, Cefriel, in qualità di UI designer con l'incarico di lavorare su IGAS.
Cefriel è un'azienda di soluzioni tecnologiche innovative con sede a Milano. Ha già avuto in passato esperienze di assessment e simulazioni d'attacco di spear phishing e ora vogliono portare competenze di interaction design e security.
Tra le altre compagnie coinvolte nella progettazione di Dogana, ci sono altre due compagnie italiane e altri diversi nomi importanti.
Dall'Italia provengono la Hewlett Packard (nel senso che è coinvolto il dipartimento italiano) e il CNIT, mentre tra le altre importanti realtà troviamo il ministero della difesa della Grecia, la compagnia dei trasporti pubblici di Bucarest e alcuni istituti universitari come l'AIT di Vienna e l'università Cattolica di Leuven.

*Le entità coinvolte in Dogana*


Durante il progetto, che illustrerò nei paragrafi successivi del capitolo, ho lavorato a stretto contatto con alcuni colleghi di Cefriel e altri esterni. In sede hanno lavorato con me Enrico, il project manager e esperto di cyber-security; Roberto, membro del security team che ha già lavorato con Enrico su altri progetti di spear phishing; Basilio in qualità di programmatore della UI.
Esternamente ho invece collaborato con Alessio, incaricato della programmazione degli strumenti di IGAS; Peter, accademico esperto di user experience; Adrien, il designer che ha preso in mano il lavoro da dove l'ho lasciato io al termine della collaborazione con Cefriel.Poiché la simulazione d'attacco non è obbligatoria, il report potrebbe riguardare solo la fase di information gathering: in questo caso si evidenzieranno dati come la percentuale dei dipendenti a rischio, il numero di informazioni sensibili presenti in rete, la probabilità di subire attacchi a seconda delle fonti degli elementi ritenuti più pericolosi, etc.

# progettazione della UI

La mia entrata in Cefriel e, di conseguenza, il mio ingresso all'interno del progetto Dogana hanno coinciso all'incirca con il momento in cui il team iniziava a lavorare concretamente sul progetto, dopo una prima fase in cui erano stati stabiliti i principi e le linee guida nei deliverable ufficiali, punti che potrebbero essere modificati in fase di realizzazione, nel caso in cui si dovessero riscontrare problemi tecnici o di policy.
Per questo la progettazione della UI di IGAS non è terminata con la fine della mia collaborazione, e cioè che probabilmente avrà bisogno di essere continuamente aggiornata per rispettare le modifiche del software; tuttavia il mio lavoro ha creato il primo nucleo di un'architettura unica che si potrà adattare facilmente a novità e implementazioni.

Dogana, almeno al momento del primo rilascio, non sarà un programma scaricabile, ma un'applicazione web. Nonostante le spiegazioni che ho ricevuto in proposito non siano state sempre chiare, sembra



che questo risolva alcuni problemi tecnici di programmazione di certe funzionalità. Dal momento che si tratta di uno strumento tecnico, la UI è stata affrontata per essere conversion driven più che experience driven.

IGAS sarà certamente la fase più corposa, non solo per la quantità di dati che dovrà gestire, ma anche per la complessità delle interazioni richieste; di conseguenza i problemi previsti alla partenza non sono pochi. Innanzitutto si prevede che ci saranno tipologie di dati molto diverse tra loro, che vanno dagli appunti di Pastebin, a stringhe di metadati, fino ai video caricati sui social; per questa ragione, gli strumenti di navigazione non possono essere progettati solo per certi tipi di media, ma devono tenere in conto questa varietà.
Il numero di singoli elementi previsto è invece elevatissimo. Anche solo per la ricerca sui social dei dipendenti basta fare un semplice calcolo: se ogni dipendente ha mediamente due social (per esempio tra Facebook, Instagram, Linkedin, Twitter, Pinterest) e un centinaio di post su ciascuno, fanno 200 elementi a dipendente. Basta un'azienda con 100 dipendenti per avere ventimila elementi totali. Tutto questo non include gli altri metodi di ricerca sui dipendenti, né tantomeno alcunché sulla compagnia. Per non parlare di business che di dipendenti ne hanno diverse migliaia. Questo problema si manifesta principalmente in due

modi: da un lato sforza il potere di calcolo della macchina e dall'altro rischia di non far emergere informazioni potenzialmente interessanti per il pentester, facendo sì che si perdano in un mare di informazioni poco rilevanti.
A causa della mole di IGAS bisognava trovare dunque un modo per velocizzare e automatizzare i processi; tuttavia un intervento umano è imprescindibile, quantomeno per avere la garanzia che non venga utilizzato del materiale proibito dal regolamento sulla privacy e dalle norme etiche del contratto, oltre che eventualmente per contribuire a taggare gli elementi con il loro contenuto semantico.
A questo scopo, i dati derivanti dalla ricerca sono stati divisi in tre categorie: "categorized data", "free text data" e "multimedia data".
La prima tipologia è fondamentalmente considerabile come metadati, dove a una stringa corrisponde necessariamente un contenuto; per questo, non è necessario un intervento umano che li classifichi poiché si sa esattamente cosa contengono. Alcuni esempi potrebbero essere "Software utilizzato: Microsoft Word" oppure "Religione: cattolica"; se in un assessment fosse proibito utilizzare dati di tipo religioso, il secondo può essere automaticamente escluso.
Nel secondo gruppo rientrano i testi, per esempio il contenuto di un post di Facebook, o un paragrafo di descrizione



dell'attività sul sito. Attraverso alcuni strumenti di text recognition è possibile interpretare un testo, o quantomeno individuarne l'argomento. Se questo tipo di valutazione fosse ritenuta sufficiente per il lavoro non è necessario un intervento umano, in caso contrario è possibile confermare o correggere l'interpretazione della macchina e aggiungere ulteriori dettagli attraverso i tag.
La terza e ultima categoria include tutti gli altri media, come suoni, video e immagini. Qui è sempre necessario un intervento umano poiché la macchina non è in grado di rilevare elementi che possono rendere inutilizzabile un contenuto.
Un ultimo problema è stato proprio quello di far coesistere l'interpretazione della macchina e l'intervento di una persona per velocizzare il più possibile una fase che rischia altrimenti di rallentare l'intero lavoro.

Ai fini di dettare le prime linee guida dell'interazione, ho considerato che il lavoro avverrà necessariamente su un personal computer, e non su dispositivi mobili, per diverse ragioni. Sarà per esempio richiesta una buona dose di potenza di calcolo, delle norme di sicurezza di un certo livello, la connessione a un server; inoltre si tratta di un tipo attività che viene svolta sul luogo di lavoro, e c'è bisogno di uno schermo grande per poter lavorare a lungo su certe quantità di materiale. Non è stato sottovalutato inoltre il

rischio più alto di sbagliare su un dispositivo touch e di dimensioni ridotte. Una volta che il sistema sarà stato rilasciato e messo alla prova per un certo periodo di tempo, sarà eventualmente possibile fare le dovute considerazioni per valutare se sia il caso di aggiungere la compatibilità per i dispositivi portatili.

Infine, per quanto riguarda l'identità del software, dal mio punto di vista il problema principale è stato il mettere ordine in un'équipe divisa in quattro team - per un totale di almeno trenta persone - più Peter, lo specialista di UX con il compito di revisionare e correggere i problemi di affordance dell'interfaccia, che è un professionista molto preparato, ma certe volte un po' accademico nei metodi.
Se al numero già di per sé elevato di persone si aggiunge che molte di queste sono programmatori web, ma non designer, ho dovuto considerare che ci sarebbero potute essere delle discontinuità nell'applicazione degli elementi, per esempio nelle dimensioni, nelle spaziature, nel peso dei caratteri. Pertanto, ho provato a progettare un sistema che riduca al minimo le discontinuità e che provi a nasconderle laddove, purtroppo, capiteranno.
Il tutto naturalmente nell'ottica di creare un ambiente chiaro e web friendly.



# workflow e feature

Al netto di quanto esaminato sulla piattaforma, sul progetto e di tutte le considerazioni avvenute tra la mia figura, quella di Peter e i programmatori di UI e funzioni, queste sono le funzionalità e le interazioni previste all'interno di Dogana. Ricordo ancora una volta che, sebbene al momento la proposta in vigore sia questa, non è detto che in futuro non possa essere parzialmente modificata.

**RUOLI**

Compagnia dei pen. tester: coloro che hanno acquistato Dogana

Compagnia messa alla prova: chi subisce l'ispezione e l'attacco

## OPERATORI DI IGAS

Esperti di policy e flagger: gli esperti del contratto; a loro spettano le valutazioni - anche etiche - circa l'appropriatezza del materiale usato

Tagger: coloro che hanno l'incarico di aggiungere dei tag di tipo semantico agli elementi

Data scientist: coloro che lavorano con i dati cercando insight e seguendo i trend della ricerca

### STRUTTURA

Data l'enorme quantità di dati prevista, l'obiettivo principale dell'interazione con la user interface di IGAS è quello di velocizzare il processo di ricerca il più possibile. Per riuscirci è necessario poter fare una scelta intelligente degli elementi da scaricare e analizzare, e con un interaction design solido che permetta di processare i dati velocemente.

IGAS si divide in due processi principali: Information Gathering Process (**IGP**) e Data Analysis Process (**DAP**). Il primo, già prototipato, punta a collezionare e indicizzare una selezione di dati sulla compagnia e, eventualmente, sui suoi impiegati, ed è l'elemento portante di IGAS. Il secondo è la rappresentazione visiva e quantitativa di questi dati, che dovrebbe trasformare quella che è una digital footprint rifinita in una digital shadow. Prima di progettare anche questa seconda parte occorrerà fare chiarezza sul tipo esatto di dati gestiti da IGAS; in ogni caso, l'output del DAP sarà un rapporto tecnico con grafici e rappresentazioni della footprint della compagnia.
Questi quattro processi si svilupperanno all'interno di quattro attività.

Data crawling (IGP):
Viene impostata una ricerca sui parametri desiderati, e la macchina inizia a scavare in cerca di informazioni. I "categorized data" vengono automaticamente classificati e spediti direttamente al DAP se rispettano le policy e il contratto, altrimenti vengono subito eliminati.
Tra le impostazioni si può anche scegliere le modalità di interpretazione automatica dei contenuti da parte della macchina. Questo aumenterà la durata del processo, riducendo però quella del tagging.

Flagging (IGP):
Questa attività consiste nell'affermare se un free text o un file multimediale rispettino le normative sulla privacy e i principi etici di pentester e azienda ispezionata. Tutti gli elementi che vengono flaggati come inappropriati vengono definitivamente cancellati insieme a tutte le informazioni associate, come i metadati ed eventuali valori semantici automaticamente rilevati.

Tagging (IGP)
Si aggiungono tag che conferiscono qualità semantiche all'oggetto. Quando è possibile, il tagger può anche aggiungere dei metadati sulla località e la data.

In un primo momento sembrava che il flagging dovesse avvenire contemporaneamente al tagging; tuttavia, constatando che la valutazione dell'appropriatezza dell'elemento va sempre fatta prima di qualsiasi altra considerazione sul suo contenuto, c'era spazio per separare le due fasi. In questo modo si evita l'errore di dimenticarsi di flaggare perdendo tempo con l'applicazione di tag che poi potrebbero essere semplicemente scartati insieme al dato; o, peggio ancora, di essere così distratti dal tagging di dimenticarsi di fare le dovute valutazioni preliminari, andando così ad accettare l'uso di materiale proibito dal contratto. In aggiunta, ma questo lo vediamo meglio più avanti, è stato possibile progettare due interazioni leggermente diverse per i due step, per una maggiore velocità di esecuzione dei compiti.

Digital footprint exploration (DAP)
Questa è la visione aggregata e anonimizzata dei contenuti taggati, presentata attraverso dei grafici preliminari, rappresentazioni visive e statistiche. Può essere gestita e modificata dal team di IGAS affinché li aiuti a comprendere lo scenario

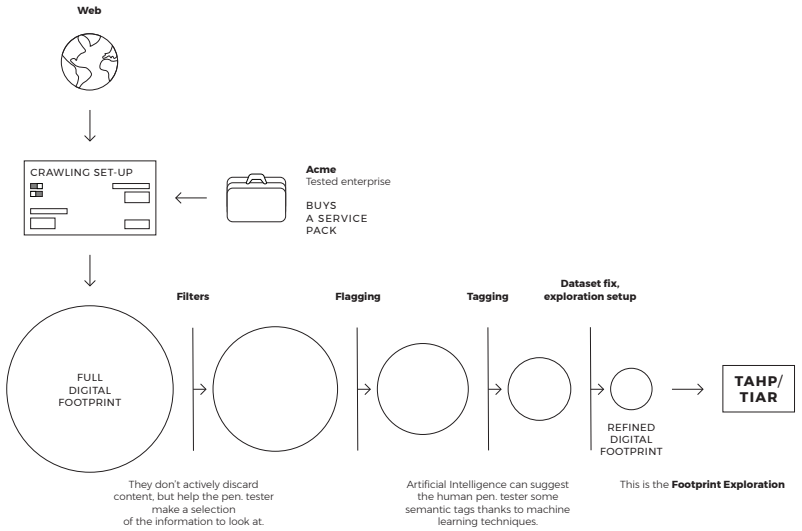


di quanto stanno ricercando.
Al contrario delle attività dell'IGP, può avere accesso alla footprint exploration anche chi è incaricato di lavorare su TAHP e TIAR (che in certi scenari, in realtà, potrebbero essere gli stessi di IGAS), ma solo dopo che IGAS l'ha validato. Inoltre TAHP e TIAR possono solamente osservare i dati, ma non modificare i grafici; nel caso in cui si rendessero conto che serve una modifica, devono interpellare IGAS.
TAHP ha bisogno di visualizzare questi dati per poter capire di più sul suo obiettivo e individuare quelli che sembrano i modi più utili di attaccare; TIAR invece lo utilizzerà come base per costruire il suo report da consegnare all'azienda alla fine della perizia.

Nonostante le attività siano logicamente consequenziali - prima di tutto si ottengono i dati, poi si scelgono quelli appropriati, li si classificano e infine li si visualizzano - è possibile che procedano in contemporanea. Per esempio è possibile che, mentre una macchina sta eseguendo la terza ricerca del progetto, un operatore stia flaggando gli elementi ottenuti dalle prime due. Il tagger può addirittura iniziare insieme al flagger: basterà che il secondo flagghi come appropriato un primo elemento che già diventa possibile passare all'attribuzione delle proprietà semantiche. Considerando che applicare dei tag richiede necessariamente più

tempo rispetto alla semplice valutazione del flagging è difficile che, lavorando in contemporanea, il flagger rimanga indietro.
Allo stesso modo, anche i data scientist non hanno bisogno che le fasi precedenti siano state completate per incominciare a produrre qualche visualizzazione. In questo modo è possibile che vedano subito dei trend, e che quindi siano in grado di dirigere la ricerca verso una precisa direzione fin da subito, anticipando i tempi e risparmiando lavoro all'IGP. C'è tuttavia da tener conto che, in base alle tecnologie che sono previste al momento in Dogana, generare una preview della footprint exploration richiede che le altre fasi siano messe temporaneamente in pausa.
Sebbene ci sia il rischio che osservare dei grafici con dati parziali, o insufficienti a costituirne un campione, sia fuorviante, è vero anche che dei data scientist siano ben consapevoli di ciò, e che quindi non incorrano in errori di questo tipo. Inoltre, la necessità di mettere in pausa il processo fa sì che il momento della generazione in itinere dei grafici venga ben ponderata.

*Flow-chart che illustra il percorso logico attraverso cui la footprint viene rifinita*

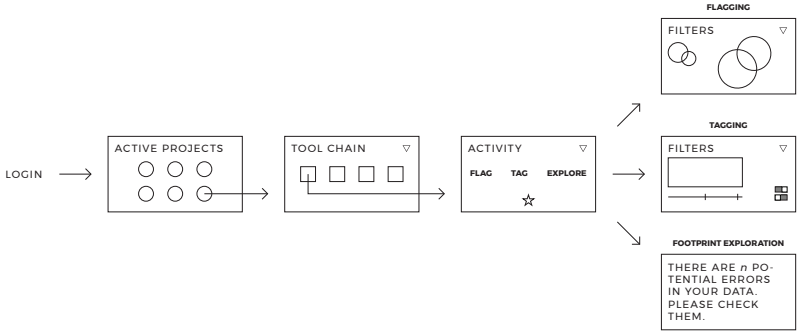


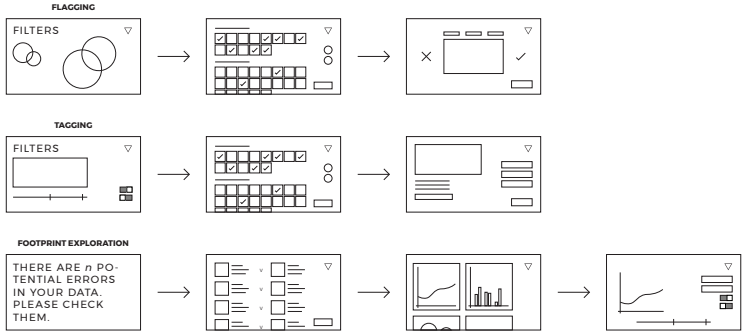



# interazione

Alla luce delle considerazioni sulla struttura e sugli obiettivi delle varie fasi segue la progettazione delle singole funzionalità.

*Flowchart dell'interazione con la UI di IGAS (parte 1)*






*Flowchart dell'interazione con la UI di IGAS (parte 2)*

## UPLOAD DI CONTENUTI RICEVUTI DALLA COMPAGNIA

Form di inserimento dei dati

Box per drag and drop: liste di dipendenti, report pre-esistenti, dataset





## SCELTA DELL'OBIETTIVO

Compagnia: non solo nell'ottica di ottenere informazioni per costruire un'esca, ma anche per trovare dettagli utili come il logo, le attività, le sedi, il numero totale di impiegati, etc.

Impiegati: può includere anche l'autocompilazione di liste e gruppi di dipendenti nel caso la compagnia non li abbia forniti

## SETUP DELLA RICERCA

Media su cui cercare: dominio, WikiLeaks, social network, etc.

Tipo di file: testo, immagini, video, audio, etc.

Massimo numero di elementi: totale della compagnia, per persona

### FILTRI

Grazie ai metadati presenti negli elementi e agli strumenti di riconoscimento automatico, i free text e i documenti multimediali ottengono subito dei tag

preliminari. In questo modo i flagger e i tagger possono selezionare subito il tipo di elementi da analizzare applicando dei filtri alla ricerca. Se da un lato questo approccio rischia di escludere dei contenuti che hanno un valore semantico ritenuto interessante dal pentester, ma che non è stato riconosciuto dalla macchina, dall'altro così facendo si è ragionevolmente sicuri di non perdere tempo visualizzando file che questo contenuto non ce l'hanno. Come al solito, la mole di dati di IGAS aiuta nella decisione: ci si aspetta di avere abbastanza dati da poter proseguire anche lasciandosi indietro gli elementi non rilevati dai filtri.

**FILTRI**

Data

Località

Tipo di file: documento, free text, video, suono, etc.

Tag semantici: indoor, figura umana, business, etc.
Possono essere aggiunti manualmente o suggeriti dalla macchina

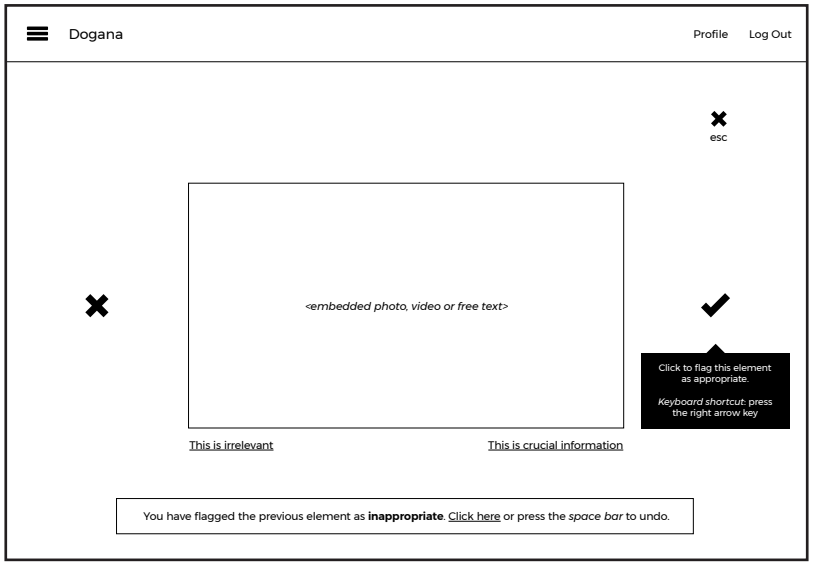




La selezione principale, e favorita dall'interfaccia, sarà tra appropriato e inappropriato. Le ulteriori due opzioni saranno disponibili ma secondarie, da scegliere solo in casi particolari - per esempio un paragrafo vuoto o una fotografia dettagliata di un badge con tanto di dettagli personali. Nel caso di un contenuto irrilevante, questo salterà la fase di tagging e finirà nelle visualizzazioni solo per i suoi metadati, ammesso che ne abbia, mentre un contenuto definito cruciale verrà visualizzato per primo dai tagger.
L'interazione che ho scelto di progettare per la maggior parte dei contenuti deriva dal caso studio di Tinder. Nonostante il fatto che probabilmente l'unico punto in comune tra un'app di incontri e lo spear phishing sia la possibilità di usare la chat per indurre le proprie conquiste nel visitare link malevoli, in realtà Tinder è un eccellente caso studio di interaction design. La possibilità di far scorrere a destra e a sinistra le foto degli altri utenti per accettare e rifiutare la conversazione consente di processare un grandissimo numero di profili in poco tempo. Inoltre si tratta di un processo di gamification, vale a dire, di trasformare un'operazione macchinosa in un gioco, aiutando a mantenere più a lungo la concentrazione.
Il contenuto, che sia un free text, un'immagine o un video, apparirà al centro

della schermata. Dal momento che le operazioni si svolgono su un computer, anziché utilizzare lo swipe ho pensato di utilizzare le frecce destra e sinistra della tasiera; rispettando la convenzione occidentale di lettura da sinistra a destra, premendo la freccia a destra il contenuto procede, dunque viene accettato, mentre premendo a sinistra torna indietro, perciò è rifiutato e cancellato dal software. Nel caso di pressione accidentale di un tasto è sempre possibile annullare l'ultima scelta premendo sulla barra spaziatrice. In questo modo è possibile eseguire l'operazione a lungo senza spostare le mani dalla tastiera.
Le tre funzioni con scorciatoia da tastiera sono comunque disponibili anche a monitor, con il pulsante per accettare sulla destra e quello per rifiutare a sinistra per aiutare anche visivamente l'intuitività della direzione da premere. Inoltre, a schermo ci sono anche i pulsanti per selezionare un file come irrilevante o cruciale, ma meno evidenti per far capire la loro natura di funzioni particolari, da usare solo in caso di effettivo bisogno.

*Il wireframe dell'interazione con cui i flagger segnalano i file che non rispettano le policy*




L'interfaccia non è applicabile ai documenti, e potrebbe non esserlo anche ad altri tipi di file al momento non ancora considerati dal team di sviluppo di Dogana. Il motivo è che alcuni formati di file, tra cui i PDF, possono portare con sé del malware; di conseguenza, per ragioni di sicurezza, non si può embeddare o visualizzare sul PC di chi acquista l'uso del framework. In questi casi, l'unica via percorribile valutata dagli sviluppatori è quella di poter assemblare degli archivi ZIP scaricabili previo un chiaro avvertimento sui rischi e chiedendo di assumersi la piena responsabilità; starà poi al flagger leggere questi documenti e valutarli singolarmente. Nota: per questo genere di file, flagging e tagging avvengono insieme per evitare di dover ripetere il procedimento due volte.

**OPZIONI E SCORCIATOIE**

Appropriato: freccia destra

Inappropriato: freccia sinistra

Annulla scelta precedente: barra spaziatrice

Appropriato, ma irrilevante: solo click

Appropriato e cruciale per la ricerca: solo click

## TAGGING

Il tagging può avvenire elemento per elemento oppure, al contrario del flagging, attraverso una selezione multipla. In questo modo si lascia maggiore discrezione al tagger, poiché può decidere di filtrare abbondantemente i risultati, così da ottenerne una selezione che gli possa interessare, e quindi valutarli nello specifico a uno a uno; in alternativa, può decidere di selezionare da un elenco più ampio tutti i file di cui riesce a riconoscere una caratteristica comune e rilevante già dall'anteprima. Anche in questo modo, come già abbiamo visto succedere altrove, si rischia di lasciare indietro qualche elemento, ma si guadagna considerevolmente sul tempo impiegato nel processo. E in effetti, anche questo già chiarito ampiamente, l'obiettivo principale di IGAS è di accelerare le operazioni



di IGP.
Per un'ulteriore velocizzazione dei meccanismi, abbiamo detto che il computer è in grado di suggerire dei tag all'utente. Sia nella visualizzazione singola che in quella multipla, l'interfaccia potrà suggerire tutti i possibili tag al pentester, il quale avrà la facoltà di confermarne la correttezza o di negarla. Sarà inoltre possibile decidere in ogni momento di accettare o rifiutare a priori tutti i suggerimenti della macchina da lì in poi. Per esempio, potrebbe scegliere di farlo se dovesse riscontrare un soddisfacente grado di affidabilità (o inaffidabilità) dello strumento, oppure in caso di un progetto dal budget ridotto in cui conta più la velocità di esecuzione che la precisione dei rilevamenti.
È possibile inoltre posticipare il tagging di un elemento in caso di dubbi.

**TAG**

Data

Località

Conferma o rifiuta i tag suggeriti dalla macchina:
un tag per volta, tutti i tag di un elemento, tutti i tag per questa ricerca

Aggiunta manuale di tag: saranno suggeriti tag simili, tag usati di recente, tag più utilizzati

## CORREZIONE DEI RISULTATI

Finalizzati i processi dell'IGP si passa alla correzione di eventuali errori, come l'unione di più profili di una stessa persona - magari perché utilizzati con email diverse, oppure perché contengono qualche dettaglio contrastante ma sono complessivamente associabili - oppure l'eliminazione di informazioni duplicate, l'unione di più tag con significato comune, etc.

### POSSIBILI ERRORI

Duplicati

Outlier: età superiore ai cent'anni e altri valori numerici improbabili

Campi lasciati vuoti o nulli

Discrepanze di informazioni

Accorpamento di tag

Fusione di profili



## ESPLORAZIONE DELLA FOOTPRINT RIFINITA

Come anticipato durante l'introduzione alla struttura del software, la funzione di digital footprint exploration non può essere ancora concretizzata poiché mancano ancora le informazioni necessarie. Tuttavia, si è ragionato su alcune delle funzioni in previsione dei prossimi step.
Il team di IGAS può utilizzare queste ed eventuali altre funzioni che saranno aggiunte in seguito per creare visualizzazioni e grafici personalizzati. Dovrà quindi farne una selezione e renderli disponibili ai team di TAHP e TIAR, che li consulteranno per portare a termine i rispettivi compiti. Potranno esplorare i dati condivisi da IGAS e aggiustarne i parametri di visualizzazione, ma non intervenire direttamente sui grafici per modificarli. Non sono inoltre abilitati ad accedere alla visualizzazione dei singoli file, ma possono solamente vederne le statistiche anonimizzate: per esempio possono vedere quanti elementi sono stati taggati con la dicitura "tempo libero", ma non le singole foto e i post.

## VISUALIZZAZIONI E FEATURE DELLA DAP

Mappa interattiva dei contenuti

Timeline interattiva

Tag cloud

Trend dei tag: nel tempo, per località, in associazione con altri tag, etc.

Momenti di maggiore attività sui social: dell'anno, della settimana, del giorno, etc.

Combinazioni personalizzate dei filtri

Salvataggio del grafico allo stato corrente

Visualizzazione di report precedenti: di precedenti assessment con Dogana oppure indipentendi

Confronta dato o grafico con quello dell'assessment Dogana precedente (ove presente)

Gruppi di dipendenti: per grado, dipartimento, gruppi di interesse, stato di rischio, etc.

# l'eredità di dogana

Dogana è uno strumento che consente di ridurre al minimo uno dei più grandi rischi in cui incorrono le aziende al giorno d'oggi: l'attacco informatico.
Non si tratta tuttavia di limitare il numero dei casi d'attacco, bensì di generare consapevolezza nelle aziende, e di conseguenza nelle persone, circa le implicazioni che può avere una gestione distratta della propria attività online.
L'eredità del software è dunque ben più grande di quel che sembra, poiché la consapevolezza che ciò che facciamo sul web ha delle conseguenze tangibili, e che nelle mani sbagliate può divenire uno strumento capace di ritorcersi contro, può dare un contributo cruciale all'eterna disputa tra i pregi e i difetti di poter condividere informazioni in tempo reale.
Può aiutare nella scelta del tipo di dati che siamo disposti a condividere, con chi e verso quali istituzioni; ci dà i mezzi per indignarci di fronte a un uso scorretto o

per noi immorale; ci guida a capire chi siano i buoni e chi i cattivi, le istituzioni che hanno a cuore l'umanità e quelle che le preferiscono il profitto.
È possibile che venga meno la natura di totale anonimato e libertà selvaggia, che una volta era dell'internet, ma d'altra parte non si può e, per tante ragioni, non si deve invertire la tendenza del progresso di una tecnologia che può dare così tanto alla nostra società. Una tecnologia per esempio capace di prevedere il collasso di infrastrutture architettoniche; in grado di rilevare con anticipo mutamenti dello stato di salute; che incentivi il progresso tecnologico con innovazioni come le automobili a guida autonoma, che si stima possano ridurre gli incidenti approssimativamente del 90%.

Da un'altra prospettiva, però, Dogana evidenzia anche una situazione piuttosto preoccupante. Dall'esperienza di Cefriel, infatti, emerge che mediamente più di una persona su tre cade in una trappola di spear phishing, cliccando un link maligno, e che la maggior parte di queste inserisce anche le proprie credenziali nel sito contraffatto che si apre. Se i numeri sono così elevati quando gli attacchi non oltrepassano i vincoli legali, etici e di privacy, c'è da aspettarsi che il dato sia addirittura una stima al ribasso: di sicuro un hacker non si fa problemi ad abusare di informazioni riservate, o di utilizzare il vero logo della compagnia che attacca.



A questo si aggiunga che, per quanto i vincoli dei test locali di Cefriel non siano così restrittivi come quelli di un progetto internazionale, gli strumenti che sono stati usati in passato non sono paragonabili a Dogana per qualità, né ampiezza di ricerca e attacco. Inoltre è evidente, una volta considerati tutti i limiti tecnologici con cui si è scontrato lo sviluppo del framework, che il livello possibile di automazione oggi è ancora troppo scarso e inaffidabile per prescindere, anche in parte, dall'intervento umano; ma, sulla base del trend dei progressi fatti nell'ambito dell'intelligenza artificiale, un giorno non troppo lontano la situazione potrebbe sbilanciarsi dall'altra parte e aumentare ancora questi dati.
Al di là della minaccia di hacking, il dato preoccupa poiché significa che nel mondo esistono delle organizzazioni che hanno in mano dei database che, nelle mani sbagliate, possono rivelarsi delle autentiche armi.

Quello che possiamo fare oggi è di utilizzare la tecnologia con consapevolezza, senza privarci di strumenti che ci semplificano o migliorano la vita, ma sempre con attenzione alle implicazioni. Il più delle volte basta chiedersi se il gioco valga la candela: se abbiamo un impegno di lavoro da non mancare e non conosciamo la zona, può valere la pena condividere la nostra posizione con Google Maps, ma forse tenerla sempre attiva per farsi ri-

cordare dove abbiamo parcheggiato non altrettanto. Donare l'accesso completo al profilo Facebook per partecipare al quiz "Che personaggio di Harry Potter sei?" certamente no.

# bibliografia:


INTERNET E LIBERTÀ

J.C.R. LICKLIDER, W.E. CLARK, On-line man computer communication, 1962, consultabile su http://www.cs.kent.edu/~javed/internetbook/hobbestimeline/HIT.html

IL WEB 2.0 APRE A SCENARI RIVOLUZIONARI

http://www.cs.kent.edu/~javed/internetbook/hobbestimeline/HIT.html

DARCY DINUCCI, Fragmented Future, in “Print”, 1999 consultabile su http://darcyd.com/fragmented_future.pdf

http://www.internetbusinessmodels.com/merchant/digitalenterprise.org/models/models.html

http://www.alexa.com/topsites

https://ifttt.com/

https://www.arduino.cc/

HANNES GRASSEGGER, MIKAEL KROGERUS, The Data That Turned the World Upside Down, in “Wired”, 28 gennaio 2017, visitabile su https://motherboard.vice.com/en_us/article/big-data-cambridge-analytica-brexit-trump

STRUMENTI E METODI
DI RACCOLTA DEI DATI

Facebook, Normativa sui Dati, aggiornata al 29 settembre 2016, consultabile su: https://www.facebook.com/about/privacy/

Marco Ronchi, lezione sul Digital Marketing, nel corso di Digital

Strategy di Design della Comunicazione, Politecnico di Milano, 5 aprile 2016.

Twitter, API Overview, consultabile su https://dev.twitter.com/overview/api

The Foundation for Data Innovation, consultabile su https://www.oracle.com/big-data/index.html

Cookie e privacy: istruzioni per l'uso, consultabile su http://garanteprivacy.it/cookie

GIUDITTA MOSCA, Perché i siti vi chiedono il consenso di usare cookie sui vostri computer, in "Wired", 25 maggio 2015, consultabile su https://www.wired.it/internet/regole/2015/05/25/cookie-law/

ALESSANDRO LONGO, Cookie Law, il chiarimento del Garante, in "Wired", 4 giugno 2015, consultabile su https://www.wired.it/internet/regole/2015/06/04/cookie-law-chiarimento-garante/

The Murky World of Third Party Web Tracking, in "MIT Technology Review", 12 settembre 2014, consultabile su www.technologyreview.com/s/530741/the-murky-world-of-third-party-web-tracking

PAUL PAPAS, How Big Data Is Revolutionizing Design, in "Wired", novembre 2014, consultabile su https://www.wired.com/insights/2014/11/how-big-data-is-revolutionizing-design/

YVES DE MONTCHEUIL, Facebook: A Decade of Big Data, in "Wired", marzo 2014, consultabile su https://www.wired.com/insights/2014/03/facebook-decade-big-data/

KUMAR SRIVASTAVA, The 'Adjacent Possible' of Big Data: What Evolution Teaches About Insights Generation, in "Wired", dicembre 2014, consultabile su https://www.wired.com/insights/2014/12/the-adjacent-possible-of-big-data/

TIM HARFORD, Big data: are we making a big mistake?, In "Financial Times", 28 marzo 2014, consultabile su https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0

GARY MARCUS, ERNEST DAVIS, Eight (No, Nine!) Problems With Big Data, in "The New York Times", 6 aprile 2014, consultabile su https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

GARY LANGER, Growing Doubts About Big Data, in "abc News" 8 aprile 2014, consultabile su http://abcnews.go.com/blogs/politics/2014/04/growing-doubts-about-big-data/

DAVID LAZER, RYAN KENNEDY, What We Can Learn From the Epic Failure of Google Flu Trends, in "Wired", 10 gennaio 2015, consultabile su https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Your Digital Footprint Matters, in "Internet Society", consultabile su http://www.internetsociety.org/your-digital-footprint-matters

Digital footprint definition, in "TechTerms", consultabile su https://techterms.com/definition/digital_footprint

https://www.digitalshadows.com/the-digital-shadow/

ATTORI

https://www.facebook.com/about/privacy

http://www.wired.com/2014/11/facebook-revamps-privacy-policy/

http://www.pcworld.com/article/2047749/facebook-s-data-policy-changes-put-your-face-front-and-center.html

https://www.washingtonpost.com/business/technology/facebook-makes-changes-to-its-data-use-policies/2013/11/15/0107eab4-4e2f-11e3-be6b-d3d28122e6d4_story.html

https://nakedsecurity.sophos.com/2015/02/02/facebooks-got-a-new-privacy-policy-and-it-plans-to-share-your-data-with-partners/

https://www.washingtonpost.com/business/technology/facebook-makes-changes-to-its-data-use-policies/2013/11/15/0107eab4-4e2f-11e3-be6b-d3d28122e6d4_story.html

https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly

https://www.google.com/policies/privacy/

http://www.bbc.com/news/technology-29381114

http://www.bbc.com/news/technology-31059874

http://edition.cnn.com/2012/02/29/tech/web/protect-privacy-google/index.html

http://www.huffingtonpost.com/2012/02/29/google-privacy-policy-changes_n_1310506.html

PATRICK HOWELL O'NEILL, The FBI nabbed another serial Deep Web pedophile, in "Daily Dot", 30 settembre 2014, consultabile su https://www.dailydot.com/crime/chris-grief-hurt2thecore-arrestor/

http://www.computerworld.com/article/2475978/encryption/snowden-at-sxsw--we-need-better-encryption-to-save-us-from-the-surveillance-state.html

http://mashable.com/2013/06/17/ad-blocker-helps-ad-industry/

http://lifehacker.com/ad-blocking-extension-ghostery-actually-sells-data-to-a-514417864

http://www.ponemon.org/about-ponemon

https://blog.mozilla.org/blog/2013/01/28/privacy-day-2013/

https://www2.idexpertscorp.com/fifth-annual-ponemon-study-on-privacy-security-incidents-of-healthcare-data



https://www-03.ibm.com/security/data-breach/

https://securityintelligence.com/cost-of-a-data-breach-2016/

https://cdt.org/about/

## RISCHI DELL'USO IMPROPRIO DI DATI

ROBERT BOOTH, Facebook reveals news feed experiment to control emotions, in "The Guardian", 20 giugno 2014, consultabile su https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds

SARAH ZHANG, Scientists Are Just as Confused About the Ethics of Big-Data Research as You, in "Wired", 20 maggio 2016, consultabile su https://www.wired.com/2016/05/scientists-just-confused-ethics-big-data-research/

ROY MARSTEN, Learning to Predict Death with Big Data, in "Wired", giugno 2014, consultabile su https://www.wired.com/insights/2014/06/learning-predict-death-big-data/

HANNES GRASSEGGER - MIKAEL KROGERUS, The Data That Turned the World Upside Down, in "Motherboard", 28 gennaio 2017, consultabile su https://motherboard.vice.com/en_us/article/big-data-cambridge-analytica-brexit-trump

Information Memorandum, del "U.S. Department Of Health And Human Services", consultabile su https://www.acf.hhs.gov/sites/default/files/cb/im1504.pdf

Risk of Insider Fraud: Second Annual Study, in "Ponemon", 28 febbraio 2013, consultabile su http://www.ponemon.org/blog/risk-of-insider-fraud-second-annual-study

https://www.ashleymadison.com/

The global fallout of the Ashley Madison hack, in "France 24", 20

agosto 2015, consultabile su http://www.france24.com/en/20150820-global-fall-out-ashley-madison-hack

BRIAN KREBS, Online Cheating Site AshleyMadison Hacked, in "Krebs on Security", 19 luglio 2015, consultabile su http://krebsonsecurity.com/2015/07/online-cheating-site-ashleymadison-hacked/

SAMUEL GIBBS, Ashley Madison condemns attack as experts say hacked database is real, in "The Guardian", 19 agosto 2015, consultabile su https://www.theguardian.com/technology/2015/aug/19/ashley-madisons-hacked-customer-files-posted-online-as-threatened-say-reports

CARMEN FISHWICK, What happened when I tried to delete my Ashley Madison account, in "The Guardian", 21 luglio 2015, consultabile su https://www.theguardian.com/technology/2015/jul/21/what-happened-tried-delete-ashley-madison-account

JOSEPH COX, Hackers Just Posted a Third Dump of Alleged Ashley Madison Data, in "Motherboard", 21 agosto 2015, consultabile su https://motherboard.vice.com/en_us/article/hackers-just-posted-a-third-dump-of-alleged-ashley-madison-data

JACOB BELTRAN, Widow addresses suicide of SAPD captain linked to Ashley Madison site, in "My San Antonio", 25 agosto 2015, consultabile su http://www.mysanantonio.com/news/local/article/Widow-addresses-suicide-of-SAPD-captain-linked-to-6465568.php

CHRIS BARANIUK, Ashley Madison: 'Suicides' over website hack, in "BBC News", 24 agosto 2015, consultabile su http://www.bbc.com/news/technology-34044506

Edward Snowden: the untold story, in "Wired", agosto 2014, consultabile su https://www.wired.com/2014/08/edward-snowden/

Spy On Me, I'd Rather Be Safe, visibile su http://www.intelligencesquaredus.org/debates/spy-me-id-rather-be-safe

HUBERT SIEBEL, Transcript: ARD interview with Edward Snowden, in "Free Edward Snowden", 23 gennaio 2014, consultabile su https://edwardsnowden.com/2014/01/27/video-ard-interview-with-edward-snowden/

BARTON GELLMAN - LAURA POITRAS, U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program, in "Washington Post", 7 giugno 2013, consultabile su https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html

ENRICO FRUMENTO - ROBERTO PURICELLI, DeepSec: Social driven vulnerability assessment, 20 novembre 2014, consultabile su https://www.slideshare.net/CEFRIEL/deepsec-social-driven-vulnerability-assessment

BIAGIO SIMONETTA, Mail-esca e allegato «malevolo» così Eye Pyramid agiva indisturbato, in "Il Sole 24 ore", 11 gennaio 2017, consultabile su http://www.ilsole24ore.com/art/tecnologie/2017-01-10/mail-esca-e-allegato-malevolo-cosi-eye-pyramid-agiva-indisturbato--224344.shtml

SAM ESMAIL, eps1.0_hellofriend.mov, in "Mr. Robot" (serie TV), 2015

## DOGANA: ALL'INTERNO DI UNA DIGITAL FOOTPRINT

Deliverable del progetto, parzialmente pubblicati su http://www.dogana-project.eu/index.php/publications/deliverables

## SITOGRAFIA AGGIUNTIVA

https://www.wired.com/insights/2014/05/big-data-mathematics-effectiveness/

https://donottrack-doc.com/en/intro/

https://www.lawserver.com/security-breach-notification

http://www.ncsl.org/research/telecommunications-and-information-technology/overview-security-breaches.aspx