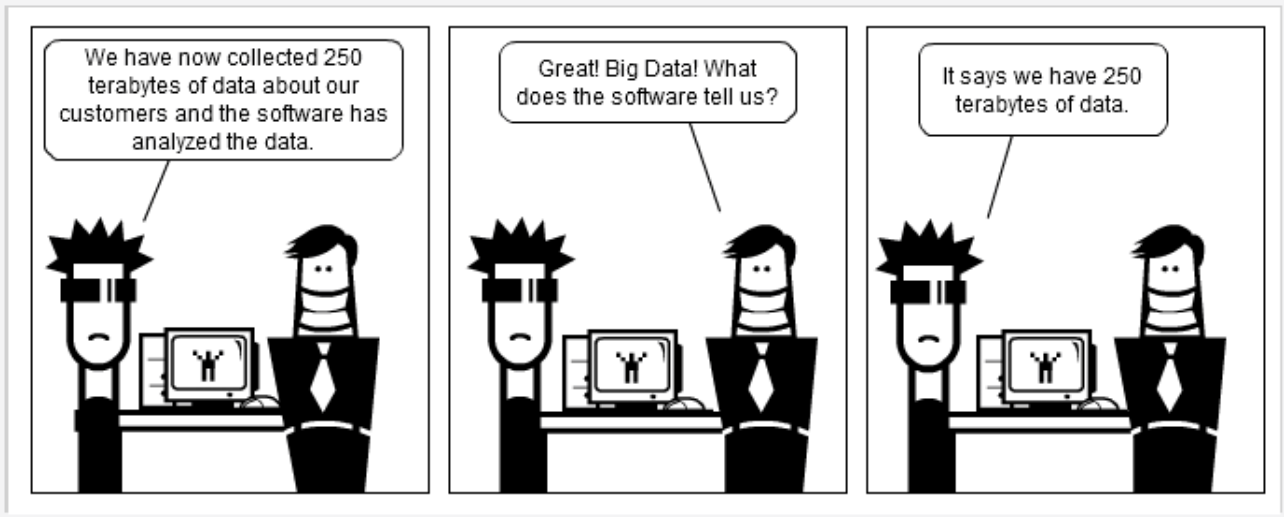
Big data

Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera n. 217/17/CONS

Servizio economico-statistico

AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI
AGCOM



"The big data Challenge"

Sean R. Nicholson
www.socmedsean.com

“Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…”

Dan Ariely,
Center for Advanced Hindsight
Duke University

"You talk to a kid these days and they have no idea what a kilobyte is. The speed things progress, we are going to need many words beyond zettabyte."

Adrian McDonald
President, Dell EMC EMEA

Servizio economico-statistico

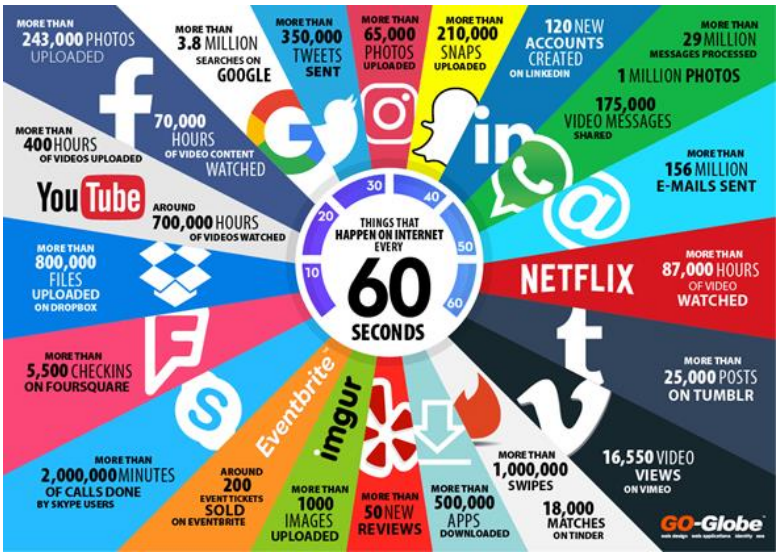
AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI
AGCOM

Anno 2018, mese di giugno

EXECUTIVE SUMMARY

Le caratteristiche dei big data

- Siamo in un'epoca in cui l'uso dei dati è oramai indispensabile nei processi decisionali di imprese, istituzioni e, sempre più, anche dei singoli cittadini. **Le attuali tecnologie consentono, infatti, la diffusione sempre maggiore dei processi di "datizzazione", un neologismo attraverso cui viene individuato quell'insieme di tecniche che consentono la conversione in formato digitale – cioè in dati** – di qualsiasi cosa (film, libri, messaggi vocali, movimenti del corpo, ecc.).
- Le parole si trasformano in dati, la posizione geografica si trasforma in dati, le interazioni sociali si trasformano in dati, anche le cose, se connesse in rete (IoT), diventano dati. Le fonti possono essere rinvenute in qualsiasi device, sensore, sistema operativo, motore di ricerca, social network.
- **L'utilizzo crescente di internet da parte degli individui, in particolare tramite i dispositivi mobili, è una sorgente inesauribile di dati**; le tracce vengono lasciate in rete in ogni momento – la cd. ***online footprint*** -, quando ci si sposta da un luogo a un altro, quando si condividono le foto o i commenti, quando si effettuano i pagamenti, quando si pratica attività sportiva, ecc..
- I *big data* rappresentano il fattore produttivo chiave in un'**economia *data-driven***; molti sono gli ambiti, sia privati che pubblici, in cui l'utilizzo di tecniche di analisi di *big data* ha permesso di creare nuovi servizi, migliorare quelli esistenti, innovare i processi produttivi e distributivi, rendere l'offerta di tutti i prodotti e servizi (anche non digitali) più rispondenti alle esigenze di consumatori e cittadini.
- I ***big data*** fanno riferimento a un salto di **paradigma interpretativo della realtà economica e sociale** attraverso tecniche di analisi (*data mining*) eseguite su **enormi quantità di dati (volume), caratterizzati da formati assai differenti (varietà), immagazzinati ed elaborati a un ritmo (velocità) sempre più rapido (spesso in tempo reale)**.

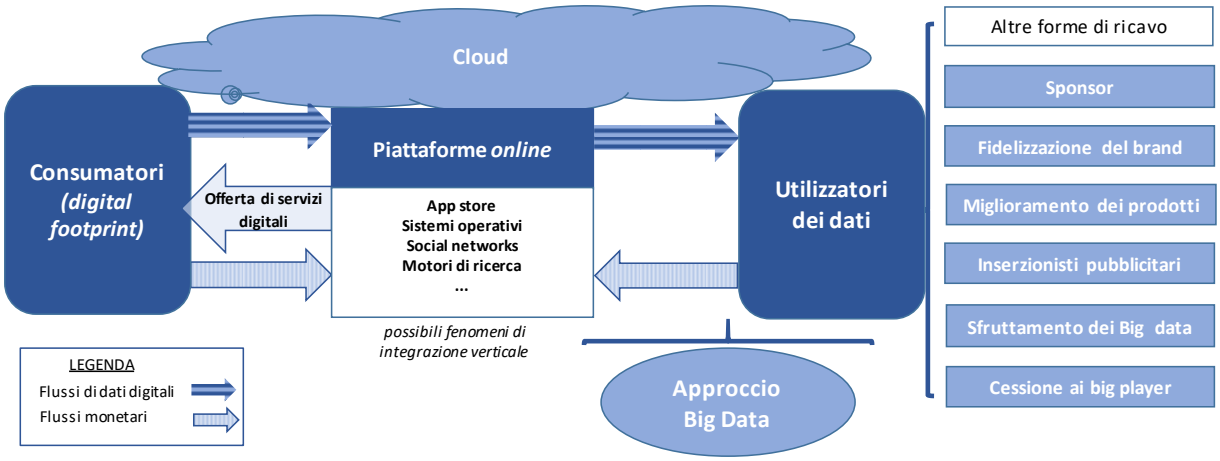


Big data: Volume, Varietà e Velocità dei dati (il flusso di dati su internet in 60 secondi)

L'ecosistema dei big data

- Nell'ecosistema dei *big data*, è possibile identificare, tra gli altri, i seguenti attori principali:
  - ✓ i **soggetti generatori di dati** (o fornitori di dati);
  - ✓ i **fornitori della strumentazione tecnologica**, tipicamente sotto forma di piattaforme per la gestione dei dati;
  - ✓ gli **utenti**, cioè coloro che utilizzano i *big data* per creare valore aggiunto;
  - ✓ i ***data brokers***, cioè le organizzazioni che raccolgono dati da una varietà di fonti sia pubbliche, sia private, e li offrono, a pagamento, ad altre organizzazioni;
  - ✓ le **imprese e le organizzazioni di ricerca**, la cui attività diventa fondamentale per lo sviluppo di nuove tecnologie, di nuovi algoritmi attraverso cui esplorare i dati ed estrarre valore;
  - ✓ gli **enti pubblici**, sia in qualità di enti regolatori dei mercati, sia con riferimento alle attività della pubblica amministrazione volte a migliorare i prodotti e i servizi offerti alla cittadinanza e in grado di aumentare il benessere collettivo.

- L'ecosistema dei *big data* presenta un **grado di interconnessione tra i vari soggetti che vi partecipano tale da rendere difficile l'identificazione di singoli mercati ben definiti**; la complessità che ne deriva determina uno scenario in cui i vari segmenti del sistema risultano spesso tra loro strettamente interrelati. Ciò conduce a un contesto in cui operano **(poche) grandi imprese multinazionali, caratterizzate da un elevato grado di integrazione in tutte le fasi dell'ecosistema, accanto a una miriade di piccole imprese specializzate**.



Rappresentazione sintetica del mercato a due versanti applicato ai dati digitali

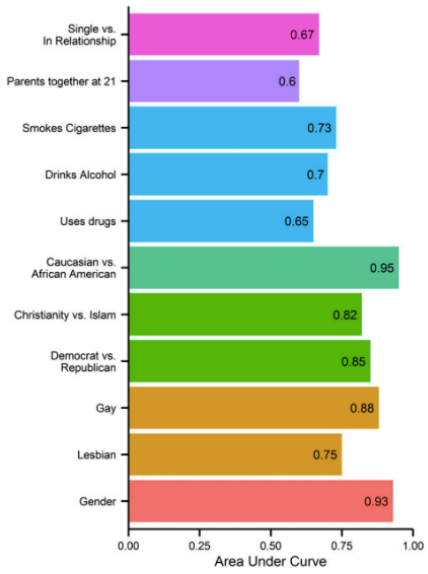
- **Fallimenti di mercato** sono legati all'esistenza di **barriere all'entrata e allo sviluppo, riscontrabili in tutte le fasi della catena del valore**.

- Uno dei segmenti principali che andrà rapidamente evolvendosi è quello relativo ai ***data center***. Al crescere della dimensione dei dati raccolti, aumenta la necessità di investire in tecnologie di acquisizione, conservazione e analisi dei dati. In questo ambito, il mercato mondiale sta convergendo verso assetti concentrati in cui spiccano le posizioni di piattaforme online quali Amazon e Google.

- Questi effetti possono presentarsi contemporaneamente in molti stadi dell'ecosistema, rafforzandosi reciprocamente e agevolando la formazione di **ambiti di mercato fortemente concentrati** (ad esempio, i mercati dei sistemi operativi, dei motori di ricerca, dei social network).

L'individuo come fonte di dati

- Ogni volta che un individuo è connesso alla rete lascia numerose "tracce", che vengono cedute agli operatori sia in modo informato, sia, più spesso, inconsapevolmente. **L'impronta digitale di ciascun individuo** si compone di numerose informazioni, alcune delle quali direttamente associabili allo stesso (nome, cognome, età, ecc.), altre associabili alle attività svolte dagli individui (pagamenti, ricerche, ecc.), altre, infine, che pur non presentando legami diretti con l'individuo, attraverso il loro processamento, possono facilmente essere associate alle persone.
- Il fenomeno dei *big data* ha reso la tradizionale **distinzione "dati personali" e "non" del tutto obsoleta** dal momento che risulta estremamente difficile stabilire *ex ante* tra tutte le informazioni raccolte su un individuo cosa rappresenta un dato personale, cosa no. Questi assumono diversa natura a seconda della quantità di dati accumulati, del contesto, nonché delle tecnologie di analisi. Ad esempio, **da un insieme, oramai anche ridotto, di dati non personali, alcune tecniche psicometriche possono facilmente derivare informazioni individuali di natura sensibile** (quali l'orientamento politico, la dipendenza da stupefacenti, ecc.).

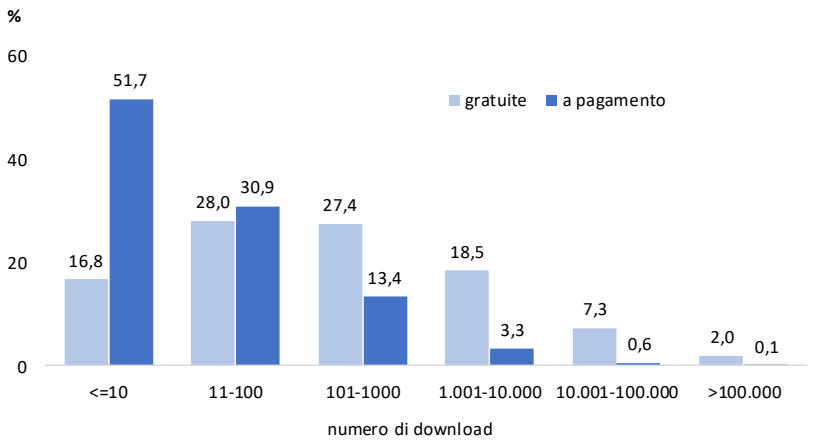


Predizioni di modelli psicometrici (risultati da 68 *like*)

- Le scelte di un individuo in ordine alla cessione di propri dati al fine di ottenere un servizio si indirizzano a seconda del bilanciamento operato tra benefici, spesso immediati (es. l'accesso a un servizio) e costi (spesso incerti e non conosciuti). In questo contesto, **l'asimmetria informativa tra utenti e operatori è pervasiva e strutturale**: non solo il consumatore non ha a disposizione tutte le informazioni di cui avrebbe bisogno per prendere una scelta informata, ma molti dei comportamenti, per essere efficienti, presupporrebbero un grado di conoscenza tecnica che va molto al di là delle competenze diffuse tra la popolazione.
- Un **maggior grado di trasparenza risulta spesso inutile laddove i consumatori non riescano, a causa di uno strutturale gap di conoscenze tecnologiche, a comprendere tali informazioni**. Scelte come quelle relative alla cessione dei propri dati, inoltre, vengono effettuate assai frequentemente di impulso e senza una valutazione delle reali conseguenze dello scambio implicito.

Lo scambio di dati: contrattazioni incomplete e mercati impliciti

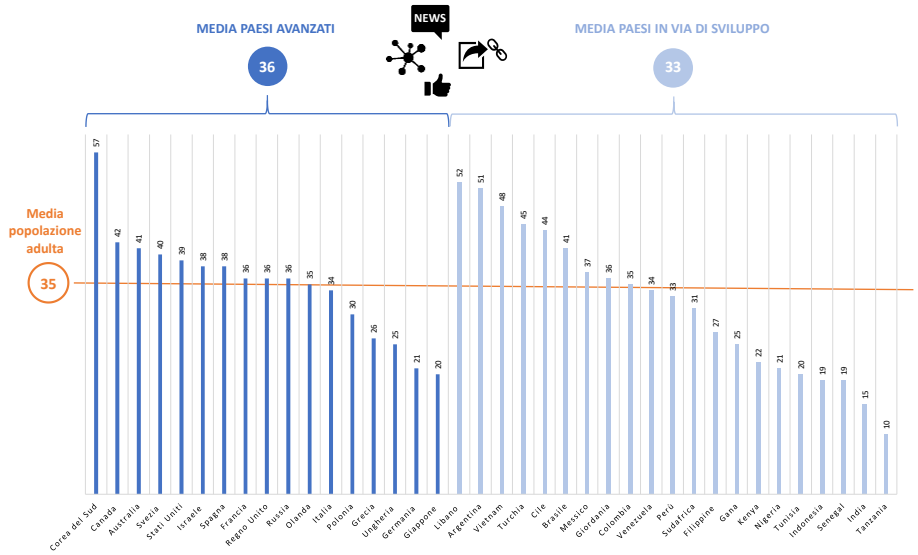
- Lo **scambio di dati dà** spesso **luogo a strutturali fallimenti di mercato**, sia perché gli investimenti posti in essere dalle imprese per la raccolta di dati sugli individui, non internalizzando i costi sociali, rischiano di condurre a un **sovrainvestimento nella raccolta delle informazioni**, sia perché, in un contesto in cui sono presenti costi di transazione e incertezza riguardo l'assegnazione dei diritti di proprietà sui dati, è probabile che le forze di mercato non siano in grado di garantire il raggiungimento di una situazione efficiente. **Si concretizza la possibilità che a prevalere siano gli interessi di coloro che detengono maggiori conoscenze tecniche e informazioni riguardo ai dati stessi**.
- Uno dei principali meccanismi attraverso cui i consumatori cedono dati digitali avviene tramite il *download* e il successivo utilizzo delle applicazioni. Gli **APP *store*** sono un importante esempio di **modalità attraverso cui vengono scambiati dati digitali.**
- Nel presente lavoro, **l'Autorità ha analizzato un dataset su oltre un milione di applicazioni**. Ne è emerso come **le APP gratuite richiedano un numero significativamente maggiore di dati individuali rispetto a quelle a pagamento**. Esiste, in sostanza, uno **scambio implicito di dati tra utenti e operatori, che si innesta nell'ambito della relazione commerciale concernente le APP.**
- L'assenza di un vero meccanismo di mercato non può che rendere queste relazioni incomplete e inefficienti. **Il consumatore non ha una chiara percezione di quali dati vengano ceduti, del loro reale valore (il prezzo) e di come gli stessi siano trattati, sia per gli usi primari, sia, a maggior ragione, per quelli secondari.** Si tratta di **una transazione *una tantum* riguardante altri beni (le APP), a fronte dell'uso dinamico dei dati degli utenti.** È, quindi, la stessa configurazione strutturale del mercato e delle relative transazioni a essere distorta e, di conseguenza, a condurre a mercati incompleti, che inevitabilmente producono risultati inefficienti e squilibrati.
- **L'andamento dei *download*, inoltre, rileva un fenomeno di "coda lunga".** Ciò determina che solo una manciata di APP, il 2%, risulta installata da un numero considerevole di utenti. **Solo 6 APP risultano installate più di 1 miliardo di volte: *Facebook*, *Google Gmail*, *Youtube*, *Google Maps*, *Google Search* e *Google Play Services*. A fronte di un numero elevatissimo di applicativi e operatori, il mercato è concentrato in poche grandi piattaforme.**



Distribuzione delle APP per numero di *download*

I big data nel sistema dell'informazione

- **L'utilizzo dei *big data* da parte di motori di ricerca e *social network* rappresenta un aspetto di particolare importanza in ragione del ruolo sempre più rilevante svolto da queste piattaforme nel sistema dell'informazione, a livello internazionale e nazionale.** Da un lato, le stesse, in virtù dei dati individuali di cui dispongono e che consentono un'accurata profilazione dell'utenza, si sono affermate come i *leader* mondiali nel settore della pubblicità online - risorsa che tuttora costituisce la fonte di finanziamento ampiamente prioritaria dell'informazione online -; dall'altro, rappresentano ormai il veicolo distributivo principale per l'informazione in rete, posto che la fruizione delle notizie su internet passa sempre più spesso attraverso questi operatori.
- **La diffusione dei *big data* sta alterando strutturalmente l'ecosistema informativo mondiale;** in particolare, i ***social network*** – in ragione del tempo trascorso dagli utenti all'interno degli stessi, delle molteplici azioni che gli individui compiono e reazioni che esprimono attraverso i propri profili/pagine/*account*, nonché delle relazioni sociali che instaurano – **si configurano certamente tra gli operatori in grado di acquisire la maggiore varietà e il maggior volume di dati sugli individui, compresi quelli relativi alle preferenze ideologiche e politiche e ai contenuti informativi letti, visualizzati, graditi, commentati e condivisi.**
- **I *social network* sono definitivamente divenuti parte integrante della dieta informativa quotidiana dei cittadini in Italia e nel mondo.**



Utilizzo dei *social network* per informarsi in Italia e nel mondo (2017; %)

- Nonostante il crescente rilievo attribuito dai cittadini ai *social network* come strumenti di informazione, sono recentemente emerse **forme patologiche** quali quelle relative alla **polarizzazione** dei cittadini (ossia la tendenza ad acquisire prevalentemente informazioni coerenti alle proprie preferenze ideologiche) e a fenomeni di **disinformazione** (quali le *fake news*).
- Per mezzo dei *social network*, i sistemi di personalizzazione automatica (che operano sulla base di algoritmi e dei *big data* acquisiti), da un lato, e le azioni di condivisione di contenuti informativi compiute dagli utenti, dall'altro, facilitano la proliferazione di notizie false e la propagazione virale di contenuti polarizzanti.

Un nuovo paradigma di policy

- Il salto tecnologico connesso all'avvento dei *big data*, e della *data-driven economy*, necessita di un **cambio di paradigma anche a livello di orientamento di policy**.
- Innanzitutto, i *big data* rendono necessario il superamento della tradizionale distinzione tra le diverse tipologie di dato (personale, sensibile, ecc.). Il **nuovo approccio deve fare riferimento al dato** *tout court*.
- **Oltre agli indiscussi benefici** economici e sociali derivanti dall'avvento della *data-driven economy*, esistono alcuni **fattori di rischio**. Si è dato ampio conto dell'esistenza di **cause di fallimento dei mercati** (quali contrattazione incompleta, mercati impliciti, asimmetrie informative, posizioni di potere di mercato). Emergono, inoltre, nuove possibili **pratiche discriminatorie**, tra le quali quelle legate al prezzo sono le più diffuse. La **discriminazione di prezzo**, che con le moderne **tecniche di profilazione online** diventa "perfetta", comporta un sicuro effetto di redistribuzione sociale appannaggio degli operatori online e, in un sistema a più versanti, a sfavore di specifiche categorie di utenti (che di volta in volta possono essere i consumatori, i lavoratori, gli editori, ecc.). Queste pratiche, anche quando sono teoricamente efficienti, **presentano rischi sociali molto significativi**. Ad esempio, la discriminazione, spesso su base algoritmica, rischia di estendersi, anche in modo involontario, a differenze nella popolazione fondate su etnia, razza, orientamento sessuale, e stato di salute.
- I **fallimenti** di mercato si ripercuotono su tutto il contesto sociale, compreso il **sistema dell'informazione**, il **pluralismo delle fonti**, e le stesse **modalità di aggregazione sociale** e di **formazione dell'opinione pubblica**.
- In conseguenza dell'esistenza di strutturali e duraturi fallimenti di mercato, è necessario, soprattutto laddove sono in discussione diritti sociali e politici, adottare un **approccio *ex ante* alla regolamentazione del dato** (e ai connessi algoritmi).
- Peraltro, questo nuovo paradigma deve considerare che le **asimmetrie informative tra utenti e operatori sono pervasive e strutturali**. In questo contesto, **è difficile ripristinare condizioni di efficienza attraverso meccanismi di trasparenza e di consenso informato**. Infatti, tali strumenti appaiono, in molti casi, insufficienti a garantire un riequilibrio conoscitivo tra operatori e consumatori, in una situazione in cui spesso soggetti quali esperti del settore, istituzioni specializzate, nonché centri di ricerca non hanno a disposizione elementi conoscitivi sufficienti a comprendere l'entità e la natura stessa dei fenomeni. In linea con quanto avviene già in contesti ad alta tecnologia (quali quelli delle comunicazioni elettroniche), appare necessario accompagnare la nuova regolazione verso **forme tecniche di regolazione diretta degli operatori che utilizzano i *big data***.
- In via preliminare, il nuovo paradigma necessita di **aprire la scatola nera** (***black box***) che regola i processi che avvengono all'interno dell'ecosistema dei *big data*, quali, tra gli altri, i **momenti e le modalità di acquisizione del dato (*data gathering & storage*), il funzionamento degli algoritmi (*algorithm accountability*), i modi di conservazione e analisi (*data analytics*), le informazioni derivate, e gli usi (primari e secondari) che ne derivano**. Rispetto a questi, e altri, aspetti si sa ancora troppo poco.
- Il **nuovo approccio deve essere, pertanto, basato su fatti, informazioni e conoscenze**. L'Autorità, in tal senso, ha già avviato ricerche con le più prestigiose università nazionali e internazionali (nel caso di questo Rapporto, il Dipartimento di Ingegneria Informatica

dell'Università "La Sapienza" di Roma) e condotto analisi con esperti del settore (nel caso della disinformazione online, con il Prof. Walter Quattrociocchi).

- Inoltre, l'Autorità ha già avviato la nuova strategia, nei suoi settori di competenza, attraverso l'**Istituzione del Tavolo Tecnico per la garanzia del pluralismo e della correttezza dell'informazione sulle piattaforme online**. Il Tavolo, sulla base dell'attuale contesto regolamentare nazionale e comunitario, cerca di declinare alcuni dei principi del nuovo approccio: aprire la scatola nera con analisi e indagini anche basate su informazioni richieste alle piattaforme online; analizzare algoritmi di *newsfeed* e raccomandazione; individuare e far emergere soluzioni collettive e condivise alle problematiche di mercato individuate; definire regole *ex ante* in capo agli operatori.

SOMMARIO

INDICE DELLE FIGURE E DELLE TABELLE

Premessa

L'Autorità, in diverse occasioni, ha evidenziato la rilevanza dei big data anche sotto il profilo del pluralismo informativo giacché il fenomeno è strettamente legato al ruolo delle piattaforme online e, quindi, all'impatto dei big data sul funzionamento dei meccanismi adottati da queste ultime nel diffondere informazione. La fruizione delle notizie in rete avviene sempre più spesso attraverso questi nuovi intermediari digitali (social network, motori di ricerca, …) che, al pari di altre piattaforme, utilizzano i dati come asset strategico, secondo la logica dei mercati multi-versante, per l'offerta di servizi e contenuti online, con la conseguente necessità di conciliare il trade-off tra il valore commerciale dell'informazione e il rispetto di diritti individuali e collettivi fondamentali quali la privacy, la tutela della concorrenza e le garanzie del pluralismo informativo.

La presenza di un fenomeno per sua natura così complesso, che si presenta caratterizzato da una forte interdipendenza e trasversalità di contenuti, e i cui effetti possono riguardare l'informazione, la concorrenza, la tutela dei consumatori e della loro privacy, presuppongono necessariamente un approfondimento delle tematiche attraverso un'indagine interdisciplinare e congiunta con l'Autorità garante della concorrenza e del mercato e il Garante per la protezione dei dati personali, quale quella avviata con la delibera n. 217/17/CONS.

Il presente Rapporto, quindi, si inserisce nell'ambito dei lavori della predetta Indagine congiunta sui big data e ne rappresenta una tappa cruciale, seppure intermedia, attraverso la quale si dà evidenza delle principali problematiche e opportunità derivanti dall'utilizzo dei big data, con particolare riferimento ai mercati (quelli delle comunicazioni) e a materie (pluralismo informativo e politico, tutela del consumatore) di stretta competenza istituzionale dell'Agcom. Questo interim report, quindi, è funzionale ai futuri lavori dell'indagine conoscitiva, rappresentando di fatto una "guida" che offre contemporaneamente strumenti, in prevalenza concetti di natura economica, e spunti per l'individuazione e la trattazione di ulteriori aspetti nelle fasi successive dell'Indagine, anche alla luce delle problematiche riscontrate.

Introduzione

In tutto il mondo, l'utilizzo innovativo dei dati nei processi decisionali sta dando vita a un radicale cambiamento, che coinvolge ogni aspetto dell'economia e della società. Questo cambio di passo, determinato dalle nuove tecnologie e tecniche per la raccolta, archiviazione e analisi dei dati, sta producendo una serie di benefici che si manifestano sia a livello individuale (consumatori e imprese) sia a livello aggregato (locale e nazionale), migliorando la qualità della vita, aprendo nuove opportunità economiche e sociali.

Secondo un rapporto di *IDC* e *Open Evidence*, il valore del mercato dei dati in Europa raggiungerà, nel 2020, i 106 miliardi di euro, a fronte di una stima pari a 60 miliardi, per il 2016, con un impatto diretto sull'intera economia continentale che raggiungerà il 4% del PIL. [1]

Il fenomeno della "datizzazione", vale a dire della trasformazione di qualsiasi informazione (film, libri, messaggi vocali, movimenti del corpo, ecc.) in dati, il progressivo aumento dell'uso di strumenti di comunicazione online da parte dei cittadini e delle imprese, nonché la conseguente crescita della digitalizzazione dei processi produttivi, non solo danno origine a un vasto ammontare di dati economici e sociali, disponibili e elaborati a una velocità sempre maggiore, ma anche a una crescente varietà di formati, ovvero, in sintesi, a quello che è stato definito come il fenomeno dei *big data*.

I *big data* rappresentano il fattore produttivo chiave in un'economia *data driven*; molti sono gli ambiti, sia privati che pubblici, in cui l'utilizzo di tecniche di analisi di *big data* ha permesso di creare nuovi servizi, migliorare quelli esistenti, innovare i processi produttivi e distributivi. Rendere l'offerta di tutti i prodotti e servizi (anche non digitali) più rispondenti alle esigenze di consumatori e cittadini.

Tale tendenza appare essere incontrovertibile e rafforzata dal fatto che, per la stragrande maggioranza degli individui, una parte rilevante della vita privata, oltre che di quella lavorativa, si è "trasferita" in rete diventando, così, una delle principali sorgenti di dati.

Nonostante i *big data* presentino delle potenzialità dirompenti, molte delle quali tuttora inesplorate, è necessario sottolineare la presenza di alcuni rischi ad essi associati. In primo luogo, l'ecosistema dei *big data* è caratterizzato dalla presenza di numerose forme di contrattazione incompleta, da mercati impliciti (ossia in cui la contrattazione del bene avviene in maniera spuria), nonché da ambiti di tipo nozionale (ossia caratterizzati da perfetta integrazione verticale e da una domanda di mercato meramente potenziale). Ciò di per sé rischia di essere fonte di severi fallimenti di mercato che pregiudicano l'efficienza sociale, statica e dinamica, dell'intero sistema. In secondo luogo, appaiono sussistere rischi di carattere collettivo, legati, tra l'altro, alla mancata incorporazione da parte del mercato di esternalità positive e negative. Tale risultato appare assai rilevante nel caso delle patologie recentemente rilevate nell'ambito dell'informazione in rete, e connesse a fenomeni quali la cd. disinformazione online.

[1] IDC e Open Vision, European Data Market SMART 2013/0063 Final Report, febbraio 2017. In questo Rapporto vengono delineati tre scenari di medio lungo periodo: uno scenario di base (o di riferimento, *Baseline scenario*), la cui ipotesi principale è il mantenimento delle attuali tendenze di crescita e sull'evoluzione delle attuali condizioni del quadro normativo europeo; il secondo scenario è quello ad alta crescita in cui il mercato dei dati entra in una traiettoria di rapida crescita (*High Growth scenario*), grazie a condizioni del quadro normativo e macroeconomiche più favorevoli; e infine uno scenario di sfida (*Challenge scenario*) in cui il mercato dei dati cresce più lentamente rispetto allo scenario di riferimento, a causa delle condizioni quadro meno favorevoli e di un contesto macroeconomico meno favorevole.

I *big data* rappresentano un ambito recente di studio e ricerca; come mostra la **Figura (i)**, che riporta l'analisi di miliardi di ricerche effettuate nel mondo dagli utenti del motore di ricerca *Google*, il termine "*big data*" presenta un *trend* fortemente crescente.[2]

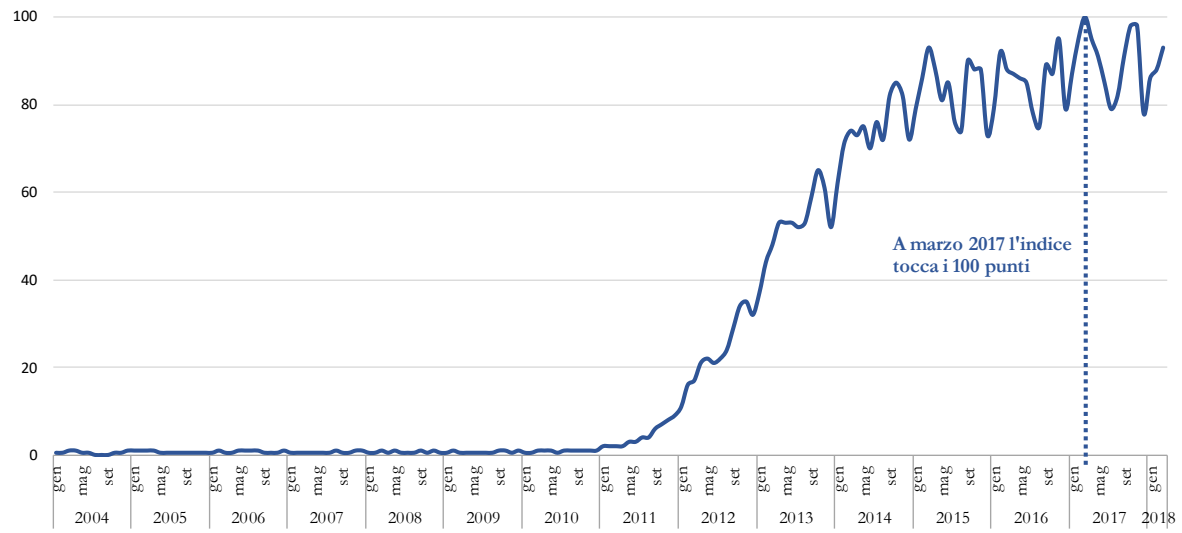


Figura (i) – *Google trends*: andamento del termine "big data" nel motore di ricerca di *Google*
Fonte: elaborazione Agcom su dati *Google.trends*

La diffusione dell'uso del termine ha, tuttavia, la conseguenza di semplificare il fenomeno. I *big data* sono, all'opposto, un fenomeno assai complesso e il loro impatto sul sistema economico e sociale deve essere valutato sulla base di una rigorosa e puntuale analisi.

Di qui la necessità di fornire a tutti gli *stakeholder* un contributo analitico che si struttura in tre parti. Pertanto, la prima parte del rapporto (Capitolo 1) mette in luce le principali caratteristiche dei *big data*, riconducibili alla crescita del volume, della varietà e della velocità con cui i dati vengono generati, acquisiti, immagazzinati e analizzati a cui si aggiunge un'analisi delle caratteristiche strutturali dei diversi ambiti di mercato che caratterizzano l'ecosistema dei *big data*. La complessità di tale ambito rende difficile delimitare con precisione i perimetri dei singoli mercati i cui confini spesso sono sovrapposti; molte aziende, inoltre, risultano integrate verticalmente o diagonalmente, e quindi presenti contemporaneamente in più segmenti. La presenza di economie di rete, di scala, di varietà, di tempo, nonché gli effetti derivanti dalle esternalità di rete, determinano strutture di mercato concentrate.

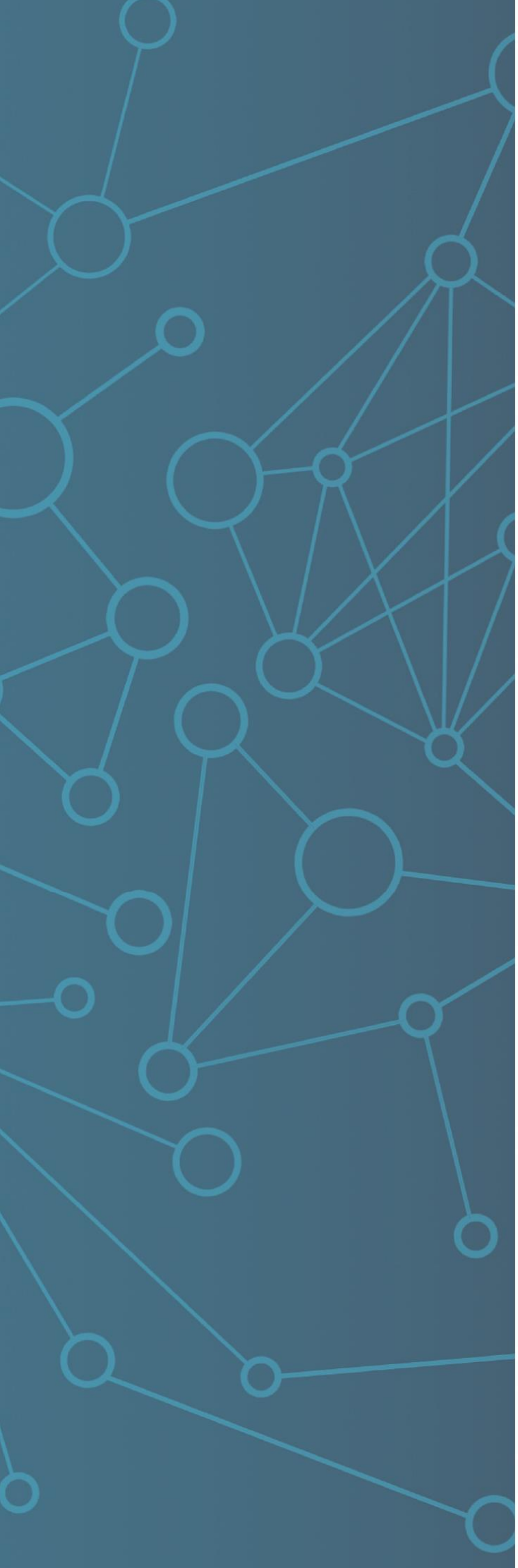
La seconda parte del rapporto (Capitolo 2) mette al centro dell'ecosistema dei *big data* il singolo individuo, ponendo particolare attenzione al superamento della tradizionale, e oramai obsoleta, distinzione tra dati personali e non. L'approccio analitico, e quindi regolamentare, deve focalizzare il proprio scopo al dato in quanto tale; ciò in virtù del fatto che è oramai impossibile individuare *ex ante* una categorizzazione dei dati: questi assumono diversa natura a seconda della quantità di dati accumulati, del contesto, nonché delle tecnologie di analisi. Ad esempio, da un insieme, oramai anche ridotto, di dati non personali, alcune tecniche possono facilmente derivare informazioni individuali di natura sensibile (quali l'orientamento politico, la dipendenza da stupefacenti, ecc.). L'individuo come fonte di dati digitali è al centro anche dell'analisi che viene proposta a conclusione del Capitolo in cui si analizza, tramite un approccio quantitativo che utilizza i *big data*, la relazione che sussiste tra l'uso sempre più massiccio di applicazioni

[2] Ogni secondo sul motore di ricerca di *Google* è possibile contare circa 67.194 interrogazioni. http://www.internetlivestats.com/one-second/#google-band

da *device* mobili e il rilascio di dati digitali. Nell'ambito di un rapporto commerciale che non appare strutturalmente ben inquadrato e codificato, ossia che stenta ad avere una struttura contrattuale ben definita, il mercato fallisce per la presenza di enormi asimmetrie informative tra i consumatori e gli operatori di servizi online. Incompletezza dei contratti che disciplinano i diritti di proprietà sui dati, assenza di espliciti mercati che regolino la formazione dei prezzi, nonché asimmetrie informative, compromettono la possibilità che il sistema converga verso un equilibrio, statico e dinamico, socialmente efficiente.

L'ultima parte (Capitolo 3) analizza gli effetti di queste problematiche relative ai *big data* sul sistema dell'informazione, e quindi sui moderni processi di formazione dell'opinione pubblica. In un contesto informativo in cui le piattaforme online, specie quelle "sociali", assumono un ruolo sempre più decisivo, i *big data* e gli algoritmi, posti a fondamento dei meccanismi attraverso i quali operano le piattaforme stesse, diventano elementi fondamentali delle democrazie avanzate. Infatti, i *big data* giungono ad avere un effetto fondamentale sul pluralismo informativo, sia dal lato della domanda, sia da quello dell'offerta.

L'obiettivo ultimo del Rapporto è, quindi, quello di ridisegnare la lettura dell'economia e della società *data – driven*, chiarendo opportunità e rischi dell'attuale contesto, in modo da favorire sia una crescita sostenuta del contesto economico sia un progresso sociale, efficiente e profondamente democratico.



1

L'ECOSISTEMA DEI BIG DATA

1.1. Le caratteristiche dei big data

Il termine *big data* fa riferimento a un nuovo approccio delle organizzazioni (imprese, enti pubblici, enti di ricerca e governi), che, tramite la combinazione di diverse banche dati e l'utilizzo di adeguati strumenti statistici e altre tecniche di *data mining*, riescono a estrarre valore dai dati. Si tratta, quindi, di **un processo di radicale riconsiderazione ed evoluzione degli approcci tradizionali all'analisi dei dati che, anche in conseguenza dei progressi tecnologici, necessitano di un nuovo paradigma interpretativo.**

Non vi è una definizione univoca del termine *big data*;[3] per la società di consulenza strategica *Gartner*, a cui molti attribuiscono un primo tentativo di definire il concetto nel 2001, i *big data* sono *"high-volume, high-velocity and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"*.[4] Con l'avvento dei *big data*, quindi, si rende necessario un nuovo approccio alla gestione dei dati per tener conto di una scala (in termini di **volume**, **velocità**) e di una complessità (**varietà**) che risulta difficile, se non impossibile, affrontare con le tecniche di analisi dei dati tradizionali.

Proprio facendo riferimento alle tre caratteristiche rilevanti dei dati - il volume, la velocità e la varietà -, è possibile mettere in evidenza questo radicale cambiamento nell'approccio all'analisi dei dati (**Figura 1.1**). Nell'epoca degli *small-data*, ossia quando i dati rappresentavano una risorsa scarsa, tipicamente era necessario porsi una domanda di ricerca e conseguentemente raccogliere i dati ("*data-is-scarce-model*"), ovvero acquisire dati su piccole parti di un universo di riferimento (un campione); le problematiche che ne scaturivano venivano trattate distintamente a seconda che si trattasse di costruire banche dati, ripetere rilevazioni di dati, ottenere risposte in tempi brevi o, anche, far "interagire" dati che presentavano formati diversi.

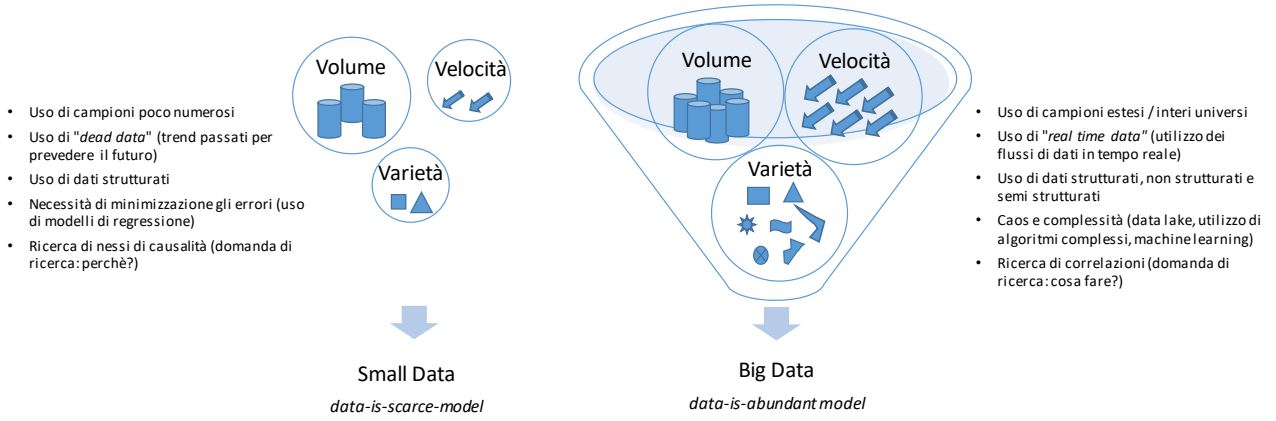


Figura 1.1 – Il cambio di paradigma con l'avvento dei big data

Fonte: elaborazione Agcom

Lo sviluppo computazionale ha consentito, con un'intensità sempre crescente, di poter affrontare la maggiore complessità di gestione delle banche dati derivante dalla sovrapposizione di almeno due delle caratteristiche sopracitate. Nell'epoca dei *big data*, i dati sono spesso raccolti a prescindere da problematiche specifiche a cui fornire una risposta: prima si raccolgono i dati, si conservano, si

[3] T. HARFORD, (2014): *Big data: are we making a big mistake?* Financial Times, 28.03.2014.

[4] Già nel 2001, la società *META Group*, in seguito divenuta *Gartner*, in un report evidenziava gli aspetti critici legati alla gestione dei dati ponendo l'attenzione su tre dimensioni; volume, velocità e varietà. D. LANEY, (2001), "*3D Data Management: controlling data Volume, Velocity and Variety*", META Group Report, File 949. Successivamente, nel 2012, in un nuovo report venne coniata la definizione. M.A. BEYER e D. LANEY, (2012), *"The importance of Big data: a Definition"*, Gartner, Analysis Report ID: G00235055.

analizzano, e, successivamente, anche in base a quello che i dati stessi comunicano, si definiscono le questioni di ricerca e commerciali (*data-is-abundant model*).

Con i *big data*, quindi, è la necessità di trattare dati che contemporaneamente presentano, con un'intensità senza precedenti, le tre succitate caratteristiche che rende le architetture tradizionali di gestione obsolete e inidonee all'analisi dei dati. Di fatto, la gestione simultanea di queste tre caratteristiche ha determinato il proliferare di tecniche di analisi di dati (*analitycs*) differenti da quelle tradizionali e tramite le quali è possibile generare valore dai *big data* sia per la risoluzione di problemi specifici, sia per l'identificazione di nuove opportunità commerciali e per la società nel suo complesso.

Quello dei *big data*, è un fenomeno dirompente, la cui portata, in termini di cambiamenti economici e sociali, non è ancora ben definita; come tutti i fenomeni di natura radicale e di portata globale, anche **i *big data* si vanno affermando portando con sé significative prospettive di crescita, economica e sociale, associate, al contempo, a dubbi e perplessità, legate al fenomeno di distruzione di mercati, imprese, mestieri e posti di lavoro, che inevitabilmente si accompagna a (e spesso precede) quello di creazione.**

Il flusso di dati attinge la sua portata dalla rapida diffusione di **forme nuove di "sorgenti dati"**; ad esempio, la diffusione dell'internet delle cose (*IoT – internet of Things*) e, più in generale, quella sempre più invasiva di sensori di ogni specie, rappresentano fonti di dati che si stanno velocemente diffondendo e integrando con una delle primarie sorgenti rappresentata dall'utilizzo massivo da parte della popolazione mondiale di telefoni cellulari di ultima generazione (v. paragrafo 2 e 0).

Per giunta, lo sviluppo vertiginoso della strumentazione tecnologica per la raccolta, conservazione, classificazione e processamento dei dati, consente alle imprese di produrre un numero sempre maggiore di informazioni, generando al contempo la necessità di avvalersi di personale specializzato, in grado di estrarre valore dai dati, e di capacità legate alla raccolta e conservazione di questa ingente mole di informazioni. Le professionalità che consentono di estrarre valore dai *big data*, quindi, rappresentano un'importantissima leva competitiva per le imprese, che hanno bisogno della disponibilità di servizi di *storage* e scalabilità dei dati. Ciò incide sulla struttura dei costi dell'ecosistema dei *big data*, e quindi sul suo grado di concorrenzialità così come su quello degli ambiti di mercato interessati da questa rivoluzione tecnologica e commerciale. Nuove sfide, dunque, emergono come conseguenza delle specifiche caratteristiche dei *big data*.

Passando all'analisi delle caratteristiche dei *big data*, le cosiddette *3V's*, è importante precisare che risulta difficile isolare le singole caratteristiche, proprio perché, come anticipato, esiste una forte interrelazione tra di esse.

1.1.1. Il volume

Il **volume** rappresenta sicuramente la caratteristica che più facilmente si può accostare ai *big data*, visto che con questo termine **si fa esplicito riferimento alla dimensione del fenomeno, ossia al grande ammontare di dati oggi disponibili, che compongono la cosiddetta *datasphere***. Numerosi studi e statistiche cercano di misurare tale caratteristica.[5] Tuttavia, a dimostrazione della difficoltà di misurare un simile fenomeno, si riscontra l'impossibilità di conoscerne con esattezza l'ammontare, data l'esistenza di stime alquanto differenti tra loro. Tutte le statistiche concordano in ogni caso su una dinamica di crescita esponenziale: un trend di sviluppo, a ritmi crescenti, che non pare volersi arrestare nei prossimi anni. Una

[5] Nell'ambito di un progetto pioneristico iniziato nel 2000, gli economisti P. Lyman e H. Varian si posero il problema di quantificare l'informazione prodotta nel mondo, con riferimento in particolare alla produzione di informazione originale; cfr. http://www2.sims.berkeley.edu/research/projects/how-much-info.

sintesi di questi risultati, che tra l'altro si collega ad un'altra caratteristica dei *big data*, la velocità, è ben catturata dalla rapidità con cui viene aggiornata l'unità di misura dell'informazione;[6] in breve tempo, infatti, l'*exabyte* si è mostrato non più idoneo a fotografare l'evoluzione del fenomeno, e si è, pertanto, passato alla misura successiva, vale a dire lo *zettabyte*, che equivale a 1000 *exabyte* (**Tabella 1.1**). In termini pratici, 1 *zettabyte* corrisponde a una capacità di archiviazione pari a oltre 36.000 anni (in termini di durata) di video in HD ovvero una pila composta da 250 miliardi di DVD.[7]

Tabella 1.1: Le unità di misura dell'informazione

Unità e simbolo	Dimensione	In termini pratici ….
Bit (b)	1 o 0	L'unità di misura elementare dell'informazione, diminutivo di "*binary digit*" che viene rappresentata alternativamente con le cifre 0 e 1, in quanto corrisponde a una scelta tra due alternative egualmente possibili.
Byte (B)	8 bits	L'informazione utile a creare un singolo carattere nel codice del computer; è l'unità di base di calcolo.
Kilobyte (KB)	1.000 o 210 bytes	Una pagina di testo equivale a 2KB. 100KB misurano una immagine fotografica in bassa risoluzione.
Megabyte (MB)	1.000KB o 220 bytes	Un file MP3 di un brano musicale "tipo" è pari a 4MB. 100MB equivalgono ad una pila di libri pari ad un metro.
Gigabytes (GB)	1.000MB o 230 bytes	Un film della durata di circa due ore può essere compresso in 1-2GB. Un testo di 1GB contiene all'incirca 1 miliardo di caratteri, ovvero circa 4.500 libri di 200 pagine o 240.000 caratteri.
Terabyte (TB)	1.000GB o 240 bytes	1TB equivale a 262.144 file MP3 (con una durata media di 4 MB). 1TB equivale a circa 4.580.000 libri di 200 pagine. Tutti i libri catalogati nella *America Library of Congress* ammontano a 15TB. Tutti i *tweet* inviati prima del 2013 equivalgono ad un file di teso di 18.5TB; per stampare un testo simile (ad una velocità di 15 pagine formato A4 per minuto) ci vorrebbero 1200 anni.
Petabyte (PB)	1.000TB o 250 bytes	1PB corrisponde a circa 4.691.000.000 libri da 200 pagine. La NSA (*National Security Agency*) analizza circa l'1,6% del traffico internet globale, circa 30PB al giorno. Volendo ascoltare 30PB di musica senza soluzione di continuità, ci vorrebbero più di 60.000 anni.
Exabyte (EB)	1.000PB o 260 bytes	1EB di dati corrisponde ad una capacità di immagazzinamento dei dati corrispondente a 33.554.432 di *iPhone 5* con una memoria di 32GB. Nel 2018, il volume mensile del traffico dati tramite telefonia mobile si stima ammonti a 1EB; se questo ammontare di dati fosse conservato in smartphone *iPhone 5* da 32 GB, sarebbe necessario formare una pila di *iPhone 5* alta 239 l'Empire State Building.
Zettabyte (ZB)	1.000EB o 270 bytes	1ZB corrisponde a 281.474.977.500.000 file MP3 della grandezza media di 4MB, ovvero 250.000.000.000 di DVD da 4,38 GB.
Yottabyte (YB)	1.000ZB o 280 bytes	Il contenuto di codice genetico appartenente ad una singola persona può essere immagazzinata in meno di 1,5GB; ciò implica che 1YB può contenere il genoma di 800 trilioni di individui, ovvero 100.000 volte circa la popolazione del pianeta.

Fonte: Elaborazione dell'Autorità su dati *Economist*, www.computerhope.com, *Cisco* e *Emmanuel Letouzé*

A livello aggregato, secondo ICD (*International Data Corporation*), si prevede, nel 2025, una massa di dati di ammontare pari a 163 *zettabyte* (**Figura 1.2**), con una crescita del volume di circa dieci volte rispetto a quello registrato nel 2016. In particolare, una mole di dati sempre maggiore deriverà dal consumo di video online e dalla presenza di sensori legati all'IoT (*internet fo Things*).

[6] Le unità di misura sono stabilite dall'*International Bureau of Weights and Measures*; in particolare, la Conferenza per i Pesi e le Misure che si è tenuta nel 1991 ha introdotto lo *Zettabyte* e lo *Yottabyte*, ad oggi l'ultima soglia conosciuta per quantificare ingenti volumi. Nel 2010, il *Product Manager* di *Google*, Jonathan Effrat, durante l'annuncio di *Google Instant*, ha dichiarato che nel web la misura di contenuti digitali nel mondo era ormai vicina allo *zettabyte*.

[7] Goodbye petabytes, hello zettabytes, https://www.theguardian.com/technology/2010/may/03/humanity-digital-output-zettabyte, The Guardian 2010;
From Bits to Brontobytes, The Oxford Math Center, http://www.oxfordmathcenter.com/drupal7/node/410;
The Zettabyte Era: Trends and Analysis, CISCO Public white paper, Giugno 2017.

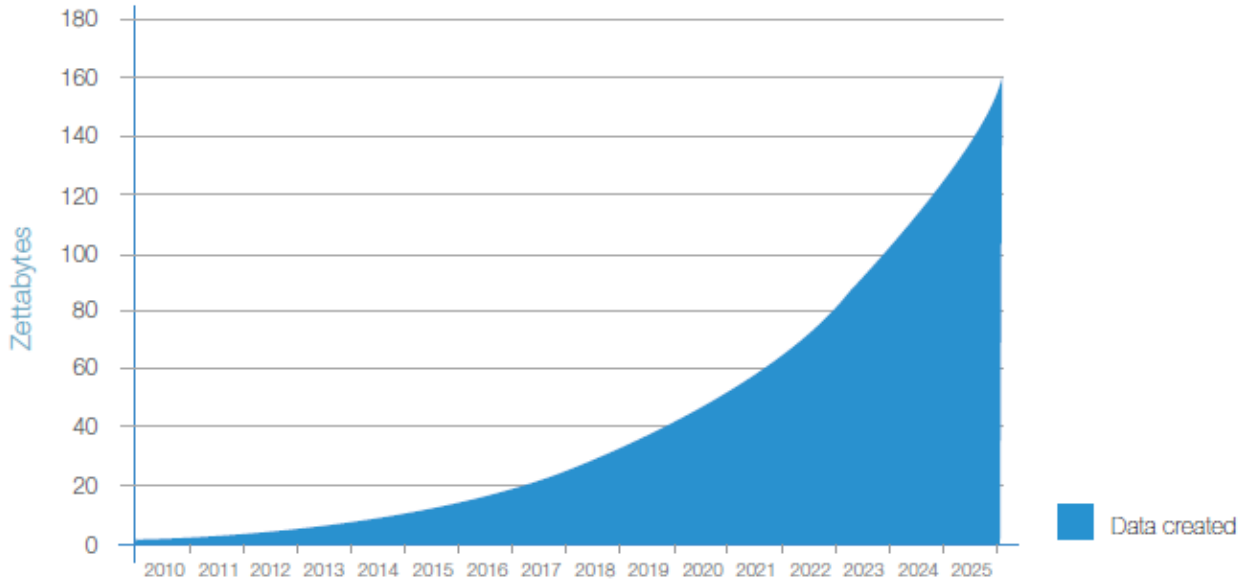


Figura 1.2 – La crescita della *datasphere* (in zettabyte)
Fonte: IDC Data Age 2025 – Aprile 2017

Va sottolineato che molti dati che vengono raccolti sono ridondanti; le stesse tecniche utilizzate nei *big data* prevedono la duplicazione dei dati al fine di preservarne la funzionalità.[8] In un approccio *big data*, la ridondanza non è infatti sinonimo di inutilità. Si pensi, ad esempio, all'insieme di informazioni che ciascun utente genera nel momento in cui svolge le proprie attività su internet (considerato come una vera e propria impronta individuale, cd. *online footprint,* cfr. paragrafo 2). Per una parte rilevante di tale ammasso di dati grezzi è stato coniato il termine di *data exhaust*:[9] si tratta di informazioni (*cookies*, file temporanei, *logfiles*, parole digitate, ecc.) che presentano al contempo un enorme volume, devono essere acquisite a grandi velocità, e sono composte dai formati più vari. Dall'analisi congiunta, e spesso in tempo reale, di questi dati è possibile estrarre un enorme valore, dal momento che vengono inferite abitudini e caratteristiche (anche sensibili) degli utenti (v. Capitolo 3 e Capitolo 5).

L'utilizzo dei *data exhaust* ha consentito, ad esempio, a *Google* di perfezionare sempre di più il proprio motore di ricerca. Più in generale, *Google* rappresenta indubbiamente un esempio di società di servizi web che ha creato un enorme valore da un "gigantesco e crescente ammasso di informazioni", molte delle quali, a prima vista, sembrano prive di importanza.

Lo sfruttamento dei dati digitali generati dagli utenti, inoltre, non solo permette di rispondere a specifiche domande di ricerca (primo utilizzo o usi primari), quando queste esistono (v. *supra*), ma anche di sfruttarne, nel tempo, il loro "valore opzionale" (utilizzi successivi o usi secondari), di cui quasi sempre non si conosce neanche l'esistenza al momento della raccolta dati. **Il valore effettivo dei dati, quindi, è nettamente superiore a quello derivante dal loro primo utilizzo**.[10] Il riutilizzo dei dati, infatti, è alla

[8] Ad esempio, il software *Hadoop*, uno dei principali strumenti che consentono di gestire diversi *petabyte* di dati, processi di elaborazione dei dati con un'alta affidabilità e la scalabilità dei dati stessi, si basa su di una riproduzione di una triplice copia (3 *factor replication*) dei *file* dati di origine, al fine di ridurre il rischio di perdita del dato stesso (cfr. paragrafo 1.2).
[9] Il termine prende spunto dalle modalità con cui questi dati vengono generati e raccolti; in maniera simile ai gas di scarico di un'auto, che fuoriescono dal tubo di scappamento (*exhaust*) collocato nella parte posteriore del veicolo, gli *exhaust data* si celano dietro le attività svolte da un utente su internet. https://whatis.techtarget.com/definition/data-exhaust
[10] Ad esempio, *Google* nel 2008 presentò per la prima volta un sistema, messo online https://www.google.org/flutrends/about/, per prevedere l'andamento dell'influenza stagionale in buona parte del mondo; l'algoritmo, sulla base delle ricerche effettuate dagli utenti sul motore di ricerca sui malanni di stagione, è in grado di realizzare una mappa aggiornata in tempo reale sulla diffusione del virus dell'influenza. Questo è un classico caso di riutilizzo di dati rappresentati dalle parole chiave digitate dagli utenti. Per completezza dell'informazione, è necessario sottolineare l'esistenza di una serie di studi che mettono in discussione l'attendibilità dell'algoritmo di *Google* che in più riprese ha sovrastimato i picchi

base dei numerosi progetti che *Google* e altre società della rete hanno in cantiere (e sono rese al pubblico spesso in versioni cd. *beta*, ossia sperimentali).

Non a caso oggi si parla di *data-driven economy* per indicare il ruolo sempre più rilevante che l'utilizzo, primario e secondario, dei dati ha nei processi decisionali dei differenti attori economici e sociali. Il crescente utilizzo della rete da parte sia dei cittadini, sia delle imprese, ha indubbiamente alimentato tale processo di crescita. I *social network*, ad esempio, hanno senza dubbio dato un forte impulso all'aumento dell'ammontare di informazioni in circolazione.[11] Si pensi, ad esempio, a *Facebook*, il *social network* più diffuso al mondo con oltre due miliardi di utenti unici; gran parte di questi utenti usa la piattaforma caricando immagini, documenti di testo, musica e video, oltre a manifestare il proprio assenso o disapprovazione. Qualsiasi passaggio/attività che l'utente effettua sul *social network* viene tracciato e si trasforma in dato.

Numerose conseguenze commerciali nascono in seguito all'aumento esponenziale del volume dei dati; innanzitutto, com'è facile immaginare, l'ammasso di dati, al fine di generare valore, deve essere archiviato (*gathering*) e conservato in maniera efficiente (*storage*). **Al crescere del volume dei dati, quindi, aumentano i costi per l'archiviazione e conservazione dei dati**, ma anche quelli legati all'estrazione di valore (*performance* dei dati) dal momento che si rende necessario l'utilizzo di sofisticati algoritmi e software e di figure professionali altamente specializzate e in grado di gestire la complessità.

1.1.2. La varietà

Il volume, come descritto in precedenza, nel mondo dei *big data* si intreccia con le altre due caratteristiche rappresentate dalla varietà e dalla velocità dei dati. La **varietà fa riferimento alla eterogeneità nelle fonti sorgenti dei dati, nei formati, con cui vengono acquisite le informazioni (tradizionali / strutturati e, soprattutto, non strutturati) e nella rappresentazione e analisi (anche semantica) dei dati immagazzinati.** L'approccio tradizionale *small data* prevede, tipicamente, l'utilizzo di dati strutturati; la gran parte dei dati, cioè, viene organizzata in strutture composte da righe e colonne che possono essere facilmente ordinate e processate secondo tecniche che fanno riferimento ai *data base* relazionali (RDBMS, fogli di calcolo, *datawarehouse*, *Costomer Relationship Management System*, ecc.), da cui è sicuramente più agevole estrarre valore anche grazie all'utilizzo di tecniche oramai consolidate. In ambito *big data*, invece, **l'eterogeneità dei dati è cresciuta in maniera esponenziale e la presenza di dati non strutturati si è andata sempre più diffondendo**; si tratta di dati che non sono organizzati secondo una precisa struttura e, di conseguenza, richiedono tecniche molto sofisticate per tramutare il dato stesso in informazione (immagini, foto, testi, email, RSS *feed*, video, sensori, *Social media*, ecc.).

Sebbene i dati strutturati siano quelli che contengono una densità di informazioni maggiore, la tendenza in atto è quella per la quale circa l'80% dei dati oggi disponibili ha natura non strutturata (**Figura 1.3**).

di influenza. D.LAZER, R. KENNEDY, G. KING e A. VESPIGNANI, (2014), *The Parable of Google Flu: Traps in Big data Analysis*, Science, Vol. 343, Issue 6176, pp. 1203-1205.

[11] Non a caso come sorgente di dati, i *social network* sono considerati come dei manicotti delle pompe dei vigili del fuoco (*fire hose data source*) o in maniera inversa, si può sostenere che *"Getting information off the internet is like taking a drink from a fire hydrant"* Mitchell David Kapor, un imprenditore Americano noto per aver sostenuto e promosso la diffusione del primo foglio di calcolo per PC, VisiCalc, e successivamente fondatore di Lotus.

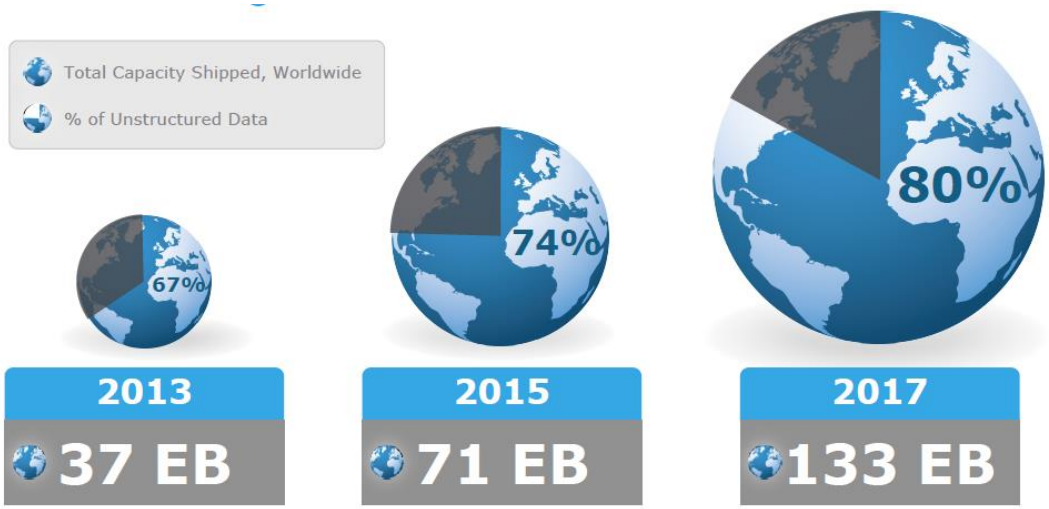


Figura 1.3 – La crescita dei dati non strutturati (in exabyte)
Fonte: IDC Structured Versus Unstructured Data: The Balance of Power Continues to Shift, March 2014

Una notevole quantità di dati viene poi definita semi-strutturati, dal momento che sebbene esistano delle possibilità di separare alcuni elementi, molta dell'informazione contenuta nel dato resta non strutturata. Il caso tipico è quello delle *email*; in effetti, qualsiasi servizio di posta elettronica presenta una serie di dati strutturati (che può essere raccolto ed organizzato in tabelle) connessi a informazioni facilmente identificabili (data e orario di invio, mittente e destinatario, ecc.). Il corpo dell'*email*, invece, generalmente è composto da un testo non strutturato, che può contenere, specie negli allegati, dati con formati assai diversi: immagini, video, audio, ecc.

Essendo i dati non strutturati la maggioranza delle informazioni oggi disponibili, è evidente che il problema della gestione dell'eterogeneità, e della sua complessità, diventa cruciale. È proprio la diversità dei formati che oggi, molto più che nell'era degli *small data*, caratterizza in maniera precipua i dati. In effetti, non tutte le imprese devono affrontare i problemi generati dalla crescita del volume e della velocità, mentre tutte, anche le più piccole, devono essere in grado di gestire la varietà dei dati, sia per la presenza di differenti sorgenti di dati, sia per le grandissime opportunità che derivano dalla possibilità di combinare dati di formato differente. Non a caso l'accesso a una grande varietà di dati – dati nuovi e vecchi, piccoli campioni e grandi campioni, dati strutturati e non strutturati, *social media data*, dati sulle scelte dei consumatori – rappresenta senza dubbio una caratteristica distintiva delle moderne piattaforme online.

Ciò è confermato dai risultati di una ricerca condotta da NVP (*NewVantage Partners*)[12] secondo la quale il 40% delle imprese intervistate – appartenenti alla classifica *Fortune 1000* stilata dalla rivista economica *Fortune* – avverte il bisogno di integrare dati con formati e fonti differenti, rispetto al 14,5% che individua nel volume e al 3,5% che indica la velocità quali fattori trainanti nelle scelte di investimento per la gestione dei *big data*.

[12] NVP, (2016), *An Update on the adoption of Big data in the Fortune 1000*.
NewVantage Partners si definisce come una società di consulenza strategica in special modo per ciò che riguarda la fissazione di una *business data strategy* http://newvantage.com/about/

1.1.3. La velocità

La **velocità** è connessa, in primo luogo, alle tempistiche con cui le banche dati vengono alimentate, in particolare **alla alta frequenza con cui i dati circolano da un punto di origine a uno di raccolta**; ciò anche in virtù della sempre maggiore disponibilità di tecnologie che consentono di raccogliere i dati in tempo reale**.** La **Figura 1.4**, fornisce una sintesi di quanti dati vengono prodotti, quindi acquisiti, su internet in soli 60 secondi, dando un'idea dell'operare congiunto delle tre caratteristiche principali dei *big data* (volume, varietà e velocità).

Tuttavia, la velocità non riguarda esclusivamente il flusso di dati, ma anche alla necessità di processare i dati in maniera rapida e per prendere decisioni ad un ritmo sempre più veloce, spesso in tempo reale (cd. *real time action*). Questo aspetto richiede competenze, infrastrutture tecnologiche, e soluzioni software di grande sofisticazione.

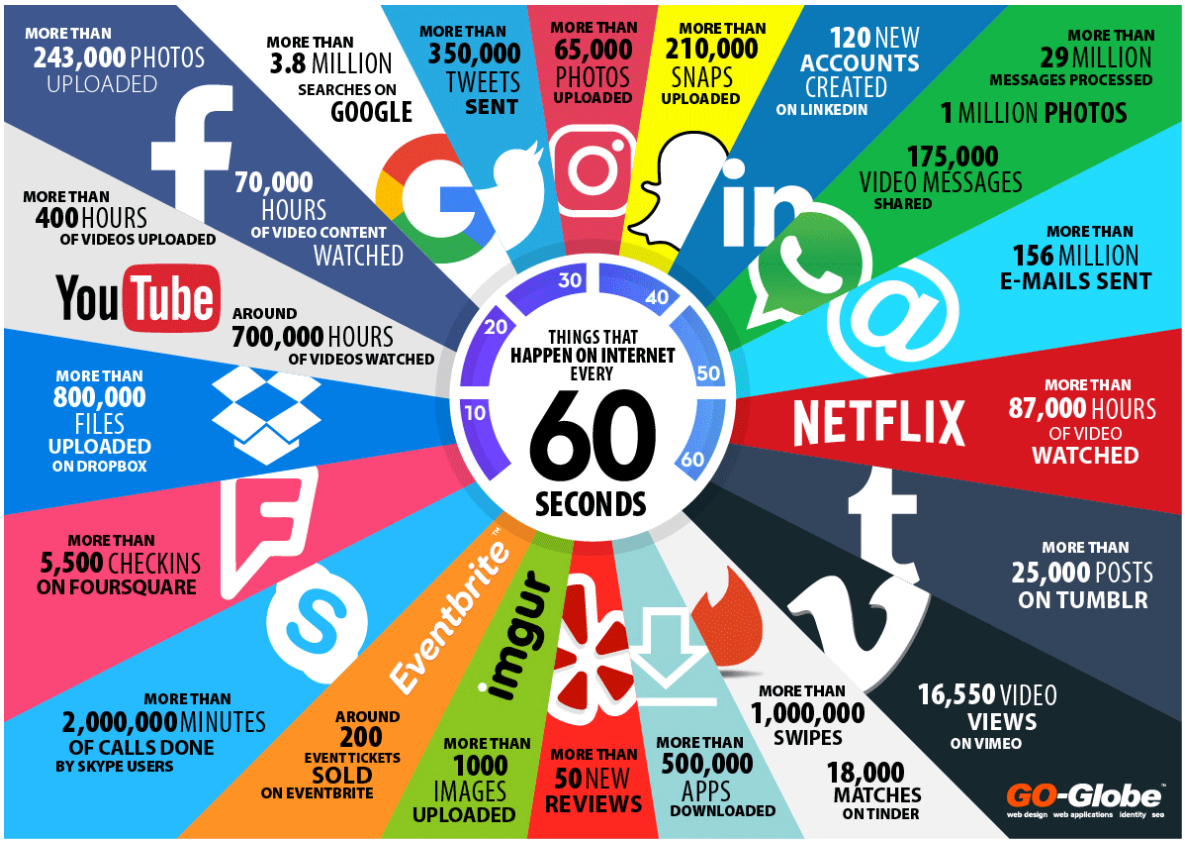


Figura 1.4 – Il flusso di dati su internet in 60 secondi
Fonte: Go-Globe.com – 2017
https://www.go-globe.com/blog/things-that-happen-every-60-seconds/

Nonostante per molte tipologie di dati resti in piedi il detto per cui "i dati sono come il vino, con l'età migliorano", volendo indicare che successivamente al suo primo utilizzo il dato non perde valore ma può essere riutilizzato per numerosi altri scopi, è altrettanto vero che molte opportunità di business sono legate alla capacità di sfruttare in maniera rapida e tempestiva i dati a disposizione. Si pensi, ad esempio, all'importanza di prendere decisioni rapide in un settore altamente competitivo come quello della vendita

al dettaglio, dove vi è l'incalzante necessità di sapere in tempo reale quali e quante tipologie di prodotti sono necessarie per rifornire efficientemente gli scaffali del negozio.[13]

La velocità dei dati ha portato a una riconsiderazione dei modelli commerciali tradizionale, tipicamente basati sull'elaborazione a blocchi dei dati (*batch processing*) connessa per lo più all'andamento passato dell'attività (cd. "*dead data*"). Tale approccio non si addice alle caratteristiche della rete. In una recente lettera agli azionisti, il fondatore di *Amazon*, Jeff Bezos, ha enfatizzato il ruolo della velocità, non solo nella gestione dei dati ma anche in quella dell'attività commerciale: "*speed matters in business*" e anche *"Most decisions should probably be made with somewhere around 70% of the information you wish you had. If you wait for 90%, in most cases, you're probably being slow."*[14] Nessuno desidera effettuare scelte sbagliate; tuttavia, temporeggiare in attesa di una "perfetta informazione", porta a ritardi decisionali e, quindi, a potenziali perdite di opportunità di *business*; sapersi correggere in corsa, sfruttando dinamicamente le informazioni dei dati, invece, può essere più efficiente nell'attuale contesto.

È importante sottolineare che l'utilizzo dei dati per prendere decisioni in tempo reale (*real-time processing*) richiede architetture software particolari che consentono di convivere con il vincolo temporale. In tal senso, come ricordato anche in precedenza, alcune tipiche attività di archiviazione, conservazione e pulizia dei dati, tipicamente svolte all'interno delle imprese, vengono sempre più spesso esternalizzate a società che si specializzano nella fornitura di questi servizi (v. paragrafo 1.5.3).

La velocità, tra l'altro, risulta molto importante anche nei processi decisionali che non riguardano attività d'impresa; in campo politico, ad esempio, lo sfruttamento dei dati è altrettanto utilizzato per cercare di ottenere un aumento del consenso, specie in periodo elettorale (v. Capitolo 3). Ma anche nel settore della sanità prendere decisioni in tempo reale può rappresentare una fonte non solo per eliminare sprechi nell'uso delle risorse, ma soprattutto per migliorare la salute degli individui, ad esempio, anticipando l'insorgere e la diffusione di malattie.

1.1.4. Le altre caratteristiche

Alle tre principali dimensioni dei *big data*, con il passare del tempo ne sono state aggiunte molte altre; ogni dimensione, rappresentata da un'ulteriore *V*, individua una caratteristica specifica dei *big data* alla quale sono associati specifici rischi e opportunità. Si è così passati in poco tempo dalle *3V*, alle *4V*, alle *7V*, fino ad arrivare, nel 2017, alle *42V's of big data*.[15] Questo processo inflazionistico, tra l'altro, non sembra arrestarsi, seppure ogni ulteriore *V* che viene individuata appare rispondere a questioni sempre più di nicchia.

Ciò nonostante, altre quattro caratteristiche dei *big data* meritano di essere citate; una fa riferimento alla capacità di estrarre **valore** economico dai *big data*.[16] **Non si tratta solo di generare valore per le**

[13] A tal proposito, un esempio di quanto la velocità sia una caratteristica estremamente rilevante nelle moderne forme di business viene dal colosso americano della Grande Distribuzione *Walmart*. Durante il periodo di Halloween, l'azienda raccoglieva dati molti positivi inerenti la vendita di uno biscotto speciale prodotto per l'occasione, tranne che in due rivendite. Da una rapida analisi dei magazzini, risultò semplicemente che tali biscotti non erano stati collocati nella maniera corretta sugli scaffali. Il fatto che un magazzino sia gestito attraverso il paradigma *Big data*, consente di conoscere in tempo reale la situazione dei singoli punti vendita e, nel momento in cui in un determinato luogo risultano vendite al di sotto di alcuni parametri, scatta un *alert* che consente un intervento rapido e mirato che, il più delle volte, consente di ripristinare le condizioni commerciali su parametri di efficienza. MARR B., (2017), *Really Big data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud*, www.forbes.com

[14] Anita Balakrishnan, *Bezos shareholder letter: Don't let the world push you into becoming a 'Day 2' company*, www.cnbc.com, 2017.

[15] T.SHAFER, (2017), *"The 42 V's of Big data and Data Science"*, Elder Research – Data Science & Predictive Analytics, https://www.elderresearch.com/company/blog/42-v-of-big-data. In questo articolo, tra l'altro, viene presentata una interessante cronologia dell'aggiunta di V alle caratteristiche dei *Big data*.

[16] S. GOGIA, (2012): *The Big Deal about Big data for customer engagement*, Forrester Research.

imprese, ma più in generale di sfruttare la crescente mole di dati per accrescere il benessere complessivo della società.

La seconda, la **veridicità**, invece, pone l'attenzione sulla rilevanza degli **aspetti qualitativi legati ai dati e, di conseguenza, alla fiducia che in essi si può riporre**.[17] Al crescere del volume, della varietà e della velocità del flusso di informazioni, ed in virtù della sempre maggiore diffusione di processi di *machine learning*, diventa cruciale per le organizzazioni lavorare con dati "credibili" affinché le analisi portino a risultati corretti.

La terza si riferisce alla **valenza** dei dati; il termine è mutuato dalla chimica per indicare il numero di elettroni che un atomo guadagna, perde o mette in comune quando forma legami con altri atomi, ossia la capacità di un atomo di creare legami. Trasportare questo concetto in ambito *big data* significa semplicemente che **più un dato è connesso con altri dati, maggiore è la sua valenza**. Due utenti *Facebook*, ad esempio, sono tra loro connessi direttamente se sono amici, così come un lavoratore è connesso al suo luogo di lavoro; i dati possono anche essere connessi in maniera indiretta, come nel caso di due scienziati che appartengono alla stessa comunità scientifica anche se non si conoscono. La valenza dei dati cresce nel tempo rendendo di fatto le connessioni fra dati sempre più dense e complesse, determinando anche in questo caso nuove sfide da affrontare.

Infine, un'ultima *V* meritevole di menzione riguarda la **visualizzazione** dei dati; **riuscire a tirare fuori informazioni sintetiche da una vastità di dati rappresenta indubbiamente una delle sfide più ardue da affrontare**. Con la giusta analisi e visualizzazione, infatti, le informazioni acquisiscono valore, altrimenti resteranno sempre a livello di dati grezzi. Tuttavia, con il termine "visualizzazione", non si fa riferimento ai classici grafici fino ad ora utilizzati (istogrammi, grafici a torte, ecc.); in effetti, si tratta di grafici complessi (infografiche) che devono avere la capacità di sintetizzare diverse informazioni senza però ridurne la portata informativa. La visualizzazione dei dati, quindi, non è complessa dal punto di vista tecnologico tanto che numerose piattaforme offrono oggi servizi per realizzare infografiche; resta, tuttavia, un'attività cruciale e complessa dal momento che riuscire a raccontare un accadimento, una storia mediante grafici risulta essere un'attività al contempo difficoltosa e essenziale.

In conclusione, le varie dimensioni dei *big data* confermano che ci si trova di fronte a un fenomeno estremamente complesso e caratterizzato da una dinamica evolutiva molto rapida; ciascuna delle caratteristiche individuate, come mostra la **Figura 1.5**, implica precise sfide da affrontare, a cui sono associati rischi e opportunità, per le imprese, i cittadini, e la società nel suo complesso.

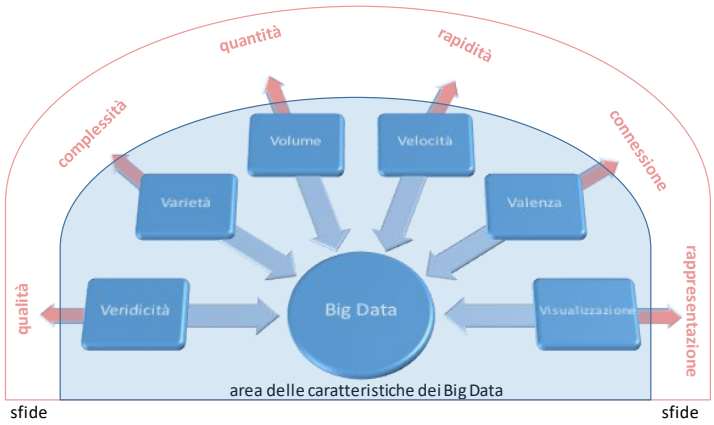


Figura 1.5 – Le caratteristiche dei *big data*
Fonte: Autorità

J. GANTZ, D. REINSEL, (2013): *The digital universe in 2020: Big data, Bigger Digital Shadows, and biggest growth in the Far East*, IDC's Digital Universe Study.

[17] M. WHITE, (2012): *Digital workplaces: Vision and reality*, Business Information Review, 29(4) 205–214.

1.1.5. Un nuovo approccio all'analisi dei fenomeni sociali

L'avvento dei *big data* determina un nuovo approccio al trattamento dei dati; la nuova filosofia, infatti, prevede che sia sufficiente ideare potenti algoritmi capaci di esplorare i dati al fine di scoprire correlazioni e regolarità (c.d. *correlation insights*).[18] Indipendentemente da qualsiasi analisi che ricerchi un nesso di causalità, come tipicamente avveniva in un'epoca *small data* (**Figura 1.1**), saranno gli algoritmi a individuare le predizioni e le azioni da intraprendere; non è più necessaria, quindi, l'individuazione di un modello comportamentale con le relative ipotesi da testare e il conseguente approccio statistico.[19] Con i *big data*, di fatto, il peso specifico delle analisi di correlazione aumenta, determinando un ribaltamento della struttura di ricerca: innanzitutto va ricercato un legame tra le variabili, solo successivamente si proverà, se necessario, a stabilire una interpretazione plausibile del fenomeno. In altre parole, sono i dati stessi a "parlare". Il superamento di un approccio basato sulla scarsità delle informazioni, e sull'analisi campionaria, ha prodotto l'affermazione delle analisi delle correlazioni dei fenomeni, con un rovesciamento epistemologico.

Tuttavia, il problema principale di un approccio basato sulle correlazioni, risiede nel fatto che all'aumentare dei dati a disposizione aumenta la probabilità di trovare variabili che tra loro presentano un legame del tutto casuale;[20] la possibilità che questi legami siano del tutto casuali, quindi, è molto elevata nei grandi *data set*, ovvero aumenta al crescere del volume e della varietà dei dati. La letteratura scientifica già da lungo tempo evidenzia il fatto che basarsi esclusivamente sulle correlazioni espone l'analisi a una serie di insidie che possono portare a interpretazioni sbagliate.[21]

Molto spesso, l'analisi di correlazione può portare a dei legami tra fenomeni che di per sé non hanno significato: se due fenomeni presentano un'elevata correlazione tra loro, non vuol dire necessariamente che tra di essi sia presente una relazione causale, potendo la correlazione derivare da un terzo fenomeno, in comune ai due analizzati, o, addirittura, essere dovuta al caso (cd. correlazione spuria), come nel caso esposto in **Figura 1.6**.

Secondo quanto riportato nella Figura, volendo portare all'estremo il concetto, si potrebbe desumere dall'analisi delle correlazioni che, ponendo in essere delle iniziative volte ad incoraggiare il consumo di mozzarella, si aumenterebbero i soggetti con una specializzazione in ingegneria civile, ovvero che, per aumentare i consumi di mozzarella, sarebbe opportuno potenziare il dottorato in ingegneria civile.

---

[18] C. ANDERSON, (2008), *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Wired*, "There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot*". È utile ricordare, a tal proposito, che la correlazione tra variabili non spiega perché due variabili si muovono in una direzione piuttosto che in un'altra, piuttosto si limita a constatare l'esistenza di un andamento. Proprio per questa ragione, la correlazione rappresenta uno strumento molto utilizzato in quanto caratterizzato da una forte potenza predittiva dal momento che consente di intervenire su di una variabile per predire (o modificare) il valore di quella, appunto, correlata.

[19] CLAUDE C.S., LONGO G., (2017), *The Deluge of Spurious Correlations in Big data*, Foundation Science, Volume 22, n. 3.

[20] GRAHAM R., SPENCER J.H., (1990), *Ramsey theory*, Scientific American, 262.

[21] FERBER R., (1956), *Are correlations any guide to predictive value?* Journal of the Royal Statistical Society Series C (Applied Statistics), 5(2).

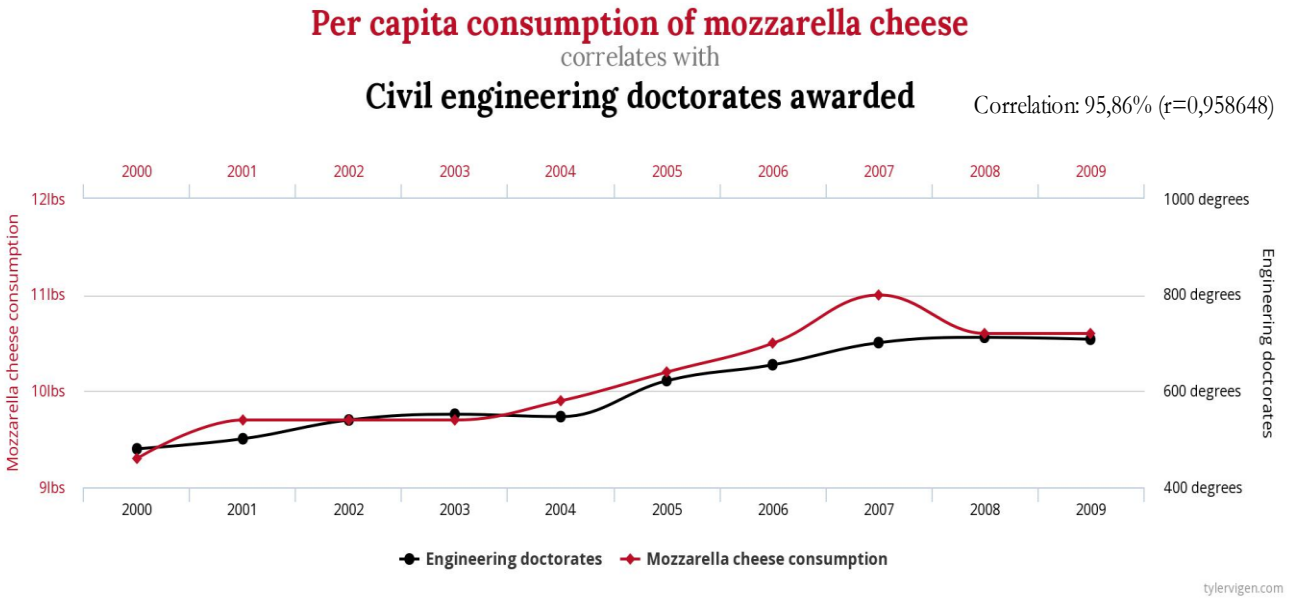


Figura 1.6 – Un caso di correlazione spuria
Fonte: http://tylervigen.com/spurious-correlations

Tuttavia, quando si tratta di prendere decisioni che riguardano le imprese, l'attenzione verso la correlazione spuria diminuisce. Per un'impresa che agisce sotto l'ipotesi di razionalità economica, infatti, ciò che conta è la possibilità di aumentare il profitto anche se tale aumento avviene tramite l'individuazione di una correlazione spuria.

In conclusione, l'avvento dei *big data* ha profondamento mutato anche l'analisi della società. Le scienze sociali in passato sono state spesso soggette a vincoli connessi alla scarsità dei dati in possesso dei ricercatori. L'avvento dei *big data* ha quindi permesso di analizzare quantitativamente temi che in passato erano lasciati ad analisi qualitative, spesso condotte con un certo grado di arbitrarietà. **La rivoluzione dei dati ha coinvolto sia la parte commerciale (il mondo delle imprese e anche delle istituzioni) sia quella scientifica, creando nuove figure professionali e un nuovo approccio all'analisi dei fenomeni. Tuttavia, alcuni dei metodi definiti in ambito *big data* necessitano di un ripensamento poiché rischiano di arrivare a conclusioni paradossali**, come quella relativa al legame esistente tra consumo di mozzarella e ingegneri civili.[22]

[22] POPPELAARS J., (2015), OR at work, https://john-poppelaars.blogspot.it/2015/04/do-numbers-really-speak-for-themselves.html

1.2. La catena del valore

Nel precedente paragrafo, è stato messo in evidenza l'elevato grado di complessità che caratterizza il mondo dei *big data*. In tal senso, l'individuazione di una **catena del valore** (o *data science process*) consente di fare un po' di chiarezza riguardo ai passaggi che seguono i dati affinché possa estrarsi valore dagli stessi. Per quanto osservato finora, infatti, **dal momento della raccolta a quello dell'utilizzo, i dati passano attraverso varie fasi tra loro interdipendenti che via via ne accrescono il valore; tali fasi possono essere assimilate ad un ciclo di vita del dato**.[23]

Occorre partire dal presupposto che **il singolo dato di per sé ha valore scarso se non nullo.** Molti ricercatori hanno accostato il ruolo che rivestono oggi i dati per l'economia e la società mondiale a quello che rappresentò il petrolio nel secolo scorso;[24] per certi versi tale analogia è calzante, soprattutto se si considera che, alla stregua del petrolio nello sviluppo dell'industria moderna, i dati oggi permettono una pletora di possibili nuovi utilizzi e quindi rappresentano un fattore produttivo determinante in un'economia basata sull'informazione. Rispetto al petrolio, tuttavia, vi sono alcune rilevanti differenze:[25] tra queste, una risiede proprio nel fatto che, mentre un barile di petrolio ha di per sé un valore, non è altrettanto vero per i dati. Quindi, **mentre per un prodotto come il petrolio la tensione tra domanda e offerta, dovuta alla presenza di una risorsa scarsa, darà origine a un prezzo di equilibrio, tale meccanismo non funziona con i dati** (cfr. paragrafo 2.7), per un esempio riguardante il settore degli applicativi mobili). Un'ulteriore **differenza risiede, come detto, nella possibilità di riutilizzo, caratteristica che i dati hanno rispetto a una materia prima come il petrolio**. Affinché i dati acquisiscano valore è poi necessaria la fase di analisi[26] la cui concretizzazione, a sua volta, non è possibile se non vengono affrontati e superati tutti i problemi generati delle caratteristiche stesse dei *big data* (v. paragrafo 1.1).

La **Figura 1.7** fornisce una possibile rappresentazione di sintesi dei processi (o macro-attività) necessari per estrarre valore dai dati.[27]

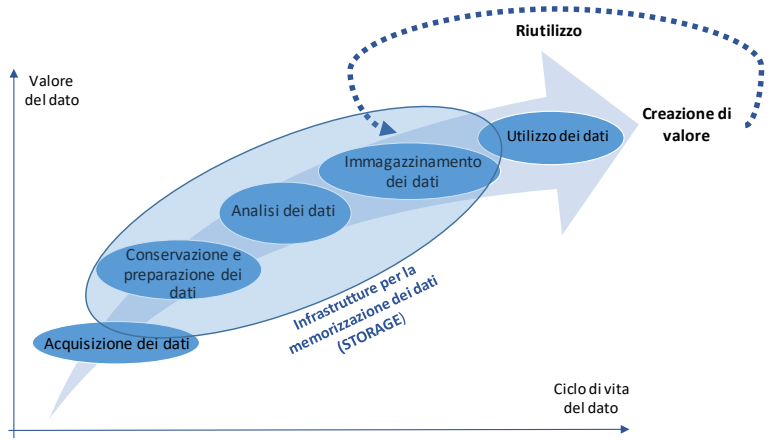


Figura 1.7 – La catena del valore nei *big data*
Fonte: Autorità

[23] FTC REPORT, (2016), *Big data: A tool for inclusion or esclusion? Understanding the issues.*
[24] The Asian Banker, (2016), *From fintech to techfin: data is the new oil*; Fortune, (2016), *Why Data is the New Oil*; Economist, (2017), *The world's most valuable resource is no longer oil, but data.*
[25] Per un maggior dettaglio sui motivi che rendono il bene dato differente dal bene petrolio, si vedano tra gli altri: J. GOLDFEIN, I. NGUYENHTTPS, (2018), *Data is not the new oil*, //techcrunch.com/2018/03/27/data-is-not-the-new-oil/; M. MANDEL, (2017), *The economic impact of Data: Why Data Is Not Like Oil*, Progressive Policy Institute.
[26] Secondo uno studio condotto nel 2013, solo lo 0,5% dei dati disponibili è analizzato. A. REGALADO, (2013), *The Data Made Me Do It*, MIT Technology Review.
[27] Per un approfondimento sul tema di veda tra l'altro il Report *European Data Market* redatto a cura della IDC e Open Evidence per conto della la Commissione Europea, febbraio 2017.

In primo luogo, c'è la loro **acquisizione**; questa fase individua **tutte le attività attraverso cui il dato viene raccolto e aggregato con altri dati e, quindi, trasportato da una sorgente a un sistema di distribuzione di dati**. Sono comprese anche tutte quelle attività svolte per controllore di quali dati già si dispone, quali sono accessibili gratuitamente, quali a pagamento, di quali strumenti si necessita per la loro raccolta, ecc. L'acquisizione dei dati di natura digitale può avvenire attraverso l'utilizzo di differenti mezzi, quali le API di chi fornisce dati provenienti dai *social network*, gli applicativi di condivisione, l'importazione di dati mediante strumenti *ETL* e l'utilizzo di strumenti di *web scraping*. In alcune circostanze, questa fase della catena del valore richiede rilevanti sforzi finanziari per gli investimenti infrastrutturali necessari, giacché sono richiesti dei requisiti minimi volti ad assicurare una bassa latenza nella fase di acquisizione dei dati e nel momento della loro interrogazione.

Vale la pena sottolineare che oggi si raccolgono dati da una varietà di fonti molto eterogenee (v. paragrafo 1.1.2). Pensiamo, ad esempio, ai *cookies* di tracciamento, ai *digital fingerprints* o alle tecniche di *history sniffing*. Attualmente una percentuale notevole di dati viene prodotta dalle attività che gli individui svolgono tramite i *device* mobili (v. paragrafo 2.2) e la produzione di dati in tempo reale sotto l'impulso delle tecnologie *IoT* (sensori). La connessione da postazione mobile, ha generato un enorme sviluppo del web in mobilità, specie nei paesi in cui sono carenti le infrastrutture di rete fissa, facendo crescere in breve tempo un ecosistema specifico in cui le aziende raccolgono dati tramite le applicazioni (cfr. paragrafo 0).

In ogni caso, l'evoluzione tecnologica ha reso disponibili strumentazioni che consentono il tracciamento "*cross-device*" delle preferenze e delle abitudini degli individui; in tal modo è possibile monitorare il singolo consumatore sfruttando dispositivi differenti, siano essi desktop, laptop, tablet, *wearable*, o smartphone. Proprio questo sistema così diffuso di raccolta dati può portare a un fenomeno definito di *over-collection*:[28] vale a dire quelle pratiche per le quali si raccolgono quantità (volume) e tipologie (varietà) di dati che vanno al di là dello scopo dichiarato.

Le aziende interessate alla raccolta dei dati puntano alla quantità e alla varietà degli stessi; tuttavia, come ampiamente ripetuto, non basta acquisire tanti e differenti dati, occorre essere in grado di analizzarli. Una volta raccolto, quindi, il dato subisce una seconda fase di lavorazione che riguarda la sua **preparazione** e **conservazione** per gli usi successivi. Si tratta di una fase in cui **il dato inizia a trasformarsi in informazione**; in questa fase volume, velocità e varietà dei dati risultano particolarmente rilevanti nella scelta riguardo alle infrastrutture tecnologiche necessarie. Affinché i dati grezzi (*raw data*) si trasformino in informazione, infatti, diventa necessario possedere le strumentazioni e le competenze che consentano di affrontare i problemi connessi alla varietà dei dati e, quindi, generare i presupposti affinché diversi tipi di dati possano interloquire tra loro, spesso in tempo reale. Si tratta di implementare quelle infrastrutture (in termini principalmente di dotazione di capacità elaborativa e software) grazie alle quali si garantisce sia una facile scalabilità dei dati, sia una loro corretta conservazione. **L'architettura che va ideata, quindi, risulta quanto mai complessa, giacché la necessità di affrontare queste criticità legate a velocità, varietà e volume dei dati richiedono tecnologie e capacità caratterizzate da una forte duttilità**.

In tal senso, nell'ottica di una maggiore integrità e fruibilità dei dati, un nuovo paradigma si è andato diffondendo, quello dei *data lake*. Il termine fu introdotto per la prima volta da James Dixon (CTO di Pentaho, società specializzata in *business intelligence*) che nel suo blog così definiva un *data lake*: "*If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a*

[28] Dal report *Big data and Privacy: A Technological Perspective* a cura del "President's Council of Advisors on Science and Technology": «*Over-collection occurs when an engineering design intentionally, and sometimes clandestinely, collects information unrelated to its stated purpose. While your smartphone could easily photograph and transmit to a third party your facial expression as you type every keystroke of a text message, or could capture all keystrokes, thereby recording text that you had deleted, these would be inefficient and unreasonable software design choices for the default text-messaging app. In that context they would be instances of over-collection*».

large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples". [29]

In estrema sintesi, i *data lake* nascono come paradigma utile a sfruttare le potenzialità generate dalle caratteristiche dei *big data* con l'intento di superare le forti rigidità presenti negli approcci tradizionali alla gestione dei dati (*data silos*, *data warehouse*, *data marts*) tipicamente incentrati sulla creazione di banche dati strutturate, quindi sicuramente facilmente interrogabili, ma gestite in maniera indipendente tra loro e caratterizzate da un pre-trattamento dei dati che spesso causava una selezione soggettiva dei soli attributi ritenuti più interessanti.[30] I *data lake*, invece, facilitano e velocizzano la condivisione dei dati proprio perché costruiti su dati grezzi – strutturati, semi-strutturati e destrutturati – nel loro formato originario e ne permettono l'analisi e, in ultima istanza, l'estrazione di valore. Il contenuto del *data lake* potrà così essere sfruttato da vari utenti che possono esaminarlo e interrogarlo alla ricerca di informazioni rilevanti (o *insight*).

Le fasi successive della catena del valore fanno riferimento a tutte quelle attività che consentono al dato di passare dallo stadio di semplice informazione a quello della conoscenza del fenomeno che si sta analizzando. Si tratta prima di tutto della fase che in **Figura 1.7** va sotto il nome di **analisi dei dati**; le attività inerenti comprendono l'esplorazione, la trasformazione e la modellazione, al fine di mettere in evidenza quelli rilevanti e, al contempo, di riuscire a sintetizzare le informazioni nell'intento, non meno importante, di portare alla luce informazioni che risultano nascoste.

La fase successiva riguarda **l'immagazzinamento dei dati**; tale processo deve rispettare dei criteri che consentono, come ripetuto più volte, una facile scalabilità, in considerazione anche del fatto che la crescita in volume dei dati ha reso sempre più sentito il problema della loro memorizzazione (cfr. paragrafo 1.5.3). **La maniera in cui ciascuna organizzazione progetta l'immagazzinamento dei propri *file*, di conseguenza, ha un rilevante impatto sulla velocità e sull'efficienza con cui avvengono l'accesso ai dati e, di riflesso, i processi decisionali**.

Per grosse aziende o per grandi moli di dati, la realtà odierna è quella dei **database distribuiti**. Nel mondo dei *big data*, i dati possono essere distribuiti sulle memorie di massa dei diversi computer (o nodi) che costituiscono la rete di un'organizzazione, i cui nodi possono essere anche fisicamente molto distanti (si pensi ad esempio alle diverse sedi di una multinazionale). Infatti, per le organizzazioni di maggiori dimensioni è impensabile mantenere tutte le informazioni in un unico punto, sia perché il traffico di accesso al singolo nodo in cui sono raccolti i dati rischia di essere eccessivo, con conseguenti problemi di congestione e quindi di ritardi nelle risposte alle interrogazioni (*query*), sia per questioni legate alla sicurezza, dal momento che concentrare i dati in un solo punto rende più vulnerabile il sistema.

[29] DIXON J., (2010), *Pentaho, Hadoop, and Data Lakes*, https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

[30] In tal senso, un classico esempio è quello dei *data silos* aziendali, ossia sistemi di gestione dei dati isolati (ad esempio un *silo* di dati per ciascuna direzione aziendale, come ad esempio i dati del libro paga, i dati finanziari, i dati dei clienti, i dati del venditore, e così via). Queste banche dati sono tipicamente costruite per scopi ben specifici, come ad esempio il recupero di un unico ordine del cliente, le elaborazioni delle buste paga alla fine di ogni mese, e così via, mentre non sono progettate per comunicare tra loro, e quindi non consentono ai singoli utenti di esplorare i dati in maniera innovativa. Tradizionalmente i dati venivano organizzati in tabelle o cartelle e file, in molti casi anche seguendo uno schema di tipo gerarchico; invece un *data lake* utilizza un'architettura piatta dove a ogni elemento che lo compone viene assegnato un identificativo univoco ed è contrassegnato con una serie di *tag* associati ai metadati del dato stesso. Un primo evidente vantaggio è l'eliminazione dei costi iniziali dell'inserimento e trasformazione dei dati stessi, ossia di tutte quelle operazioni che servono a trasformare i dati nel passaggio che va dalla sorgente al *silo*. Attraverso il paradigma del *data lake*, invece, i dati raccolti su di un determinato consumatore possono essere agganciati a una serie di altri dati provenienti da altre fonti (come ad esempio dati sul traffico, sul clima, ecc.) che seppure non risultano avere un collegamento diretto con il consumatore stesso, possono essere utilizzati in una loro combinazione al fine di estrarre più informazioni nel processo di analisi.

Tuttavia, **l'efficacia di un sistema di file distribuiti dipende dalla sua corretta gestione integrata, grazie alla quale si garantisce a tutti gli utenti del sistema informativo di un'organizzazione, qualsiasi sia la loro posizione geografica, la diponibilità di dati sempre aggiornati**; diventa quindi cruciale un'integrazione di tipo logico degli archivi anche se i dati da un punto di vista fisico possono essere distanti tra loro. La mancata integrazione, infatti, può generare problemi di duplicazione (ad esempio la stessa variabile può essere denominata in maniera diversa da nodo a nodo del sistema di file distribuito) e rischi di anomalie nell'aggiornamento, che si verificano quando uno stesso dato è aggiornato in un archivio del sistema ma non in un altro.

Tutte queste fasi della catena del valore sono propedeutiche a quella in cui i dati vengono utilizzati per assumere decisioni, stadio finale in cui il dato da semplice conoscenza si trasforma in una visione dei fatti (o *wisdom*).

La necessità di trattare grandi volumi e varietà di dati, sempre più spesso in tempo reale, genera una complessità fino ad ora sconosciuta che, di fatto, impone alle varie organizzazioni di dotarsi di infrastrutture adeguate, dal momento che l'utilizzo dei sistemi tradizionali, come i sistemi software di gestione di basi dati (DBMS) usualmente caratterizzati da un'architettura incentrata sull'utilizzo di una sola componente hardware dedicata non appaiono più funzionali.

L'ultima fase della catena del valore riguarda, dunque, l'utilizzo dei dati a supporto dei **processi decisionali**; tali attività si sostanziano, in sintesi, nella necessità di trovare una forma di raccordo tra i dati e le condotte intraprese dall'organizzazione. L'utilizzo dei dati nei processi decisionali può riguardare la riduzione dei costi di produzione, l'organizzazione del personale, l'invenzione di nuovi servizi e o prodotti, così come qualsiasi apporto al miglioramento degli indicatori di *performance*.

È importante sottolineare che a beneficiare dell'utilizzo dei *big data* non è solo il settore privato, ma anche quello pubblico, sia in termini di miglioramenti di efficienza nell'uso delle risorse, sia nella creazione di nuovi servizi e nel miglioramento di quelli attuali.

Le fasi della catena del valore presentano un certo grado di sovrapposizione dovuto principalmente al fatto che tutte devono svolgersi nel minor tempo possibile (a volte, qualche milionesimo di secondo, come nel caso delle aste per la vendita di pubblicità online).[31]

In conclusione, lo strumento della catena del valore consente di modellare il sistema dei *big data* e, di conseguenza, di identificare i vari passaggi attraverso cui generare valore, e, più in generale, conoscenza dai dati.

[31] Una delle principali infrastrutture tecnologiche attualmente utilizzate da un numero sempre crescente di imprese è quella che va sotto il nome di *Apache Hadoop*. *Hadoop* si presenta come un vero e proprio ecosistema composto da una serie di strumenti grazie ai quali è possibile processare i *big data*, ovvero navigare in un *data lake*. La complessità di tale tecnologia è tale che una sua dettagliata descrizione esula chiaramente dall'obiettivo di questo lavoro, tuttavia è interessante notare come l'impalcatura dell'architettura di *Hadoop* si poggia sulla grande capacità di elaborare rapidamente una grossa mole di dati attraverso il *Hadoop Distributed File System* (*HDFS*), ossia un sistema di archiviazione dei dati distribuito, un con un elevatissimo livello di flessibilità, tramite il quale gestire dati strutturati e non, provenienti da differenti fonti. Inoltre, *Hadoop* è costruito in maniera tale da prevedere anche una tolleranza all'errore (*fault tolerance*) per cui, nel caso in cui un nodo presenti un fallimento hardware, l'architettura è tale da reindirizzare l'attività su di un altro nodo che possiede i dati copia in maniera tale da dare continuità al processo computazionale. Il meccanismo di replicazione dei dati, inoltre, serve anche per recuperare i dati in maniera più efficiente. In conclusione, si tratta, di uno strumento tecnologico al cui interno sono ricondotte tutte le principali caratteristiche della catena del valore passate in rassegna in questo paragrafo e che per le sue prerogative rappresenta oggi uno standard utilizzato dalle principali imprese per navigare i *big data*.

1.3. I soggetti attivi

La complessità sottostante la catena del valore determina uno scenario di mercato dei *big data* molto variegato e articolato. **Se gli attori che partecipano al mercato possono essere individuati, anche se non sempre con facilità, molto più difficile risulta sciogliere l'intricato intreccio di interazioni che avvengono nel mondo dei *big data*.**

Per quanto finora descritto, nell'ecosistema dei *big data*, è possibile identificare, tra gli altri, i seguenti attori principali:[32]

a) i **soggetti generatori di dati** (o fornitori di dati);
b) i **fornitori della strumentazione tecnologica**, tipicamente sotto forma di piattaforme per la gestione dei dati;
c) gli **utenti**, cioè coloro che utilizzano i *big data* per creare valore aggiunto;
d) i ***data brokers***, cioè le organizzazioni che raccolgono dati da una varietà di fonti sia pubbliche, sia private, e li offrono, a pagamento, ad altre organizzazioni;
e) le **imprese e le organizzazioni di ricerca**, la cui attività diventa fondamentale per lo sviluppo di nuove tecnologie, di nuovi algoritmi attraverso cui esplorare i dati ed estrarne valore;
f) gli **enti pubblici**, sia in qualità di enti regolatori dei mercati, sia con riferimento alle attività della pubblica amministrazione volte a migliorare i prodotti e i servizi offerti alla cittadinanza e in grado di aumentare il benessere collettivo.

Tuttavia, l'ecosistema dei *big data* presenta un **grado di interconnessione tra i vari soggetti che vi partecipano tale da rendere difficile l'identificazione di singoli mercati ben definiti**; la complessità che ne deriva, di conseguenza, determina uno scenario in cui i vari segmenti del sistema, di cui la **Figura 1.8** offre una possibile rappresentazione, risultano spesso tra loro strettamente interrelati. Ciò determina un assetto di mercato in cui operano **(poche) grandi imprese multinazionali, caratterizzate da un elevato grado di integrazione verticale, diagonale e orizzontale in tutte (o quasi tutte) le fasi dell'ecosistema, accanto a una miriade di piccole imprese specializzate** che spesso, dopo il periodo di *start-up*, vengono acquisite da quelle più grandi.

Gli assetti competitivi negli specifici ambiti di mercato risultano strettamente dipendenti da alcune comuni caratteristiche strutturali; tale circostanza giustifica un approccio, in prima battuta, di portata più ampia, teso ad individuare una serie di caratteristiche strutturali dei *big data*, per poi approfondire gli effetti della loro concretizzazione in alcuni specifici ambiti.

Vale inoltre rilevare che, come si illustrerà in maggior dettaglio nel Capitolo 4, **l'ecosistema dei *big data* è caratterizzato dalla presenza di numerose forme di contrattazione incompleta, da mercati impliciti (ossia in cui la contrattazione del bene avviene in maniera spuria), nonché da ambiti di tipo nozionale (ossia caratterizzati da perfetta integrazione verticale e da una domanda di mercato meramente potenziale).**

Ciò di per sé è già fonte di **fallimenti di mercato che pregiudicano l'efficienza sociale, statica e dinamica, dell'intero ecosistema dei *big data***.

[32] CURRY E., (2016), *The Big data value chain: definitions, concepts and theoretical approaches*, in New Horizons for a Data-Driven Economy, Springer, Cham.

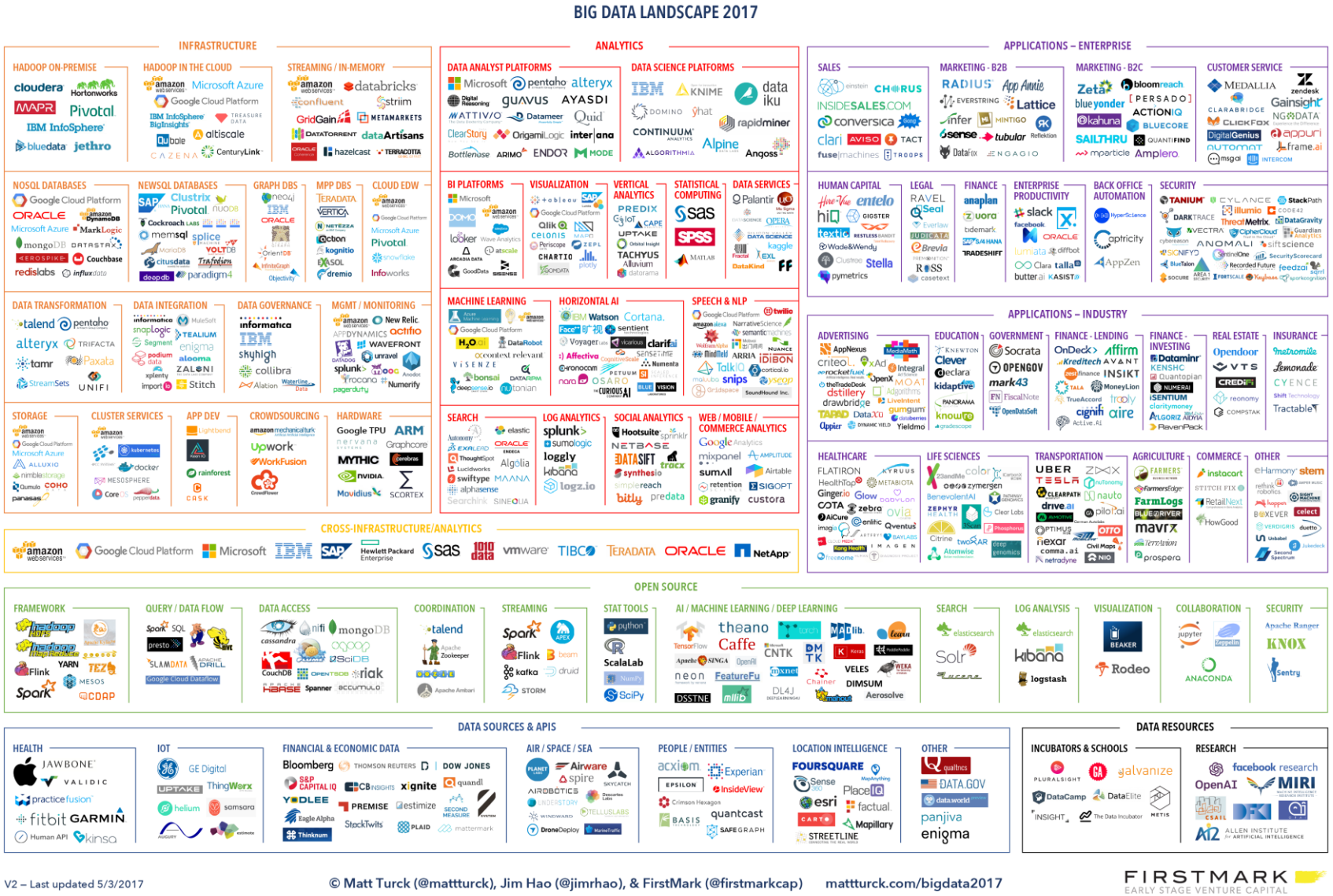


Figura 1.8 – Gli scenari di mercato nei *big data*

Fonte: http://mattturck.com/wp-content/uploads/2017/05/Matt-Turck-FirstMark-2017-big-Data-Landscape.png

1.4. Le principali caratteristiche dei mercati dei big data

Una prima fondamentale caratteristica strutturale applicabile all'ecosistema dei *big data*, in particolare alla componente dei dati digitali legati alle singole persone, è la possibilità di inquadrare tale contesto nell'ambito della teoria economica dei **mercati a due versanti (*two-sided market*) o a più versanti (*multi-sided market*)**, come verrà approfondito in seguito analizzando il caso delle *APP* (v. Paragrafo 2.5).

In generale, i mercati a due (o più) versanti si contraddistinguono *i)* per la presenza di due (o più) gruppi distinti e separati di agenti economici, le cui interazioni sono mediate da una cd. "piattaforma"; *ii)* per la stretta interdipendenza delle scelte effettuate su un versante rispetto a quelle compiute dagli agenti che operano nell'altro (o negli altri) versante.

Nel caso dei dati digitali legati alle singole persone, come si evince anche dalla sintetica rappresentazione della **Figura 1.9**, è possibile individuare nelle piattaforme online che offrono servizi ai consumatori il ruolo di intermediario tra i primi e gli utilizzatori di dati.

In primo luogo, **le piattaforme online assumono quindi un connotato tecnico di "piattaforma" nel senso della teoria a più versanti**, ossia di intermediario tra agenti economici che si situano in ambiti di mercato distinti e che "comunicano" attraverso la loro presenza.

In secondo luogo, **nel versante dei consumatori, le transazioni tra consumatori e piattaforme online sono caratterizzate da profonde e incontrastabili asimmetrie informative e da una forte e strutturale incompletezza delle transazioni, al punto che il dato digitale scambiato non assume un valore specifico (il prezzo) come in tutti i mercati, ma viene ceduto in forma spuria e non contrattualizzata**. Il paragrafo 2.7.2 è destinato a trattare nel dettaglio questo tipo di relazione e le conseguenze sull'efficienza economica dell'intero sistema, mentre, nel Capitolo 3, la prospettiva verrà ulteriormente allargata agli aspetti di natura politica e sociale.

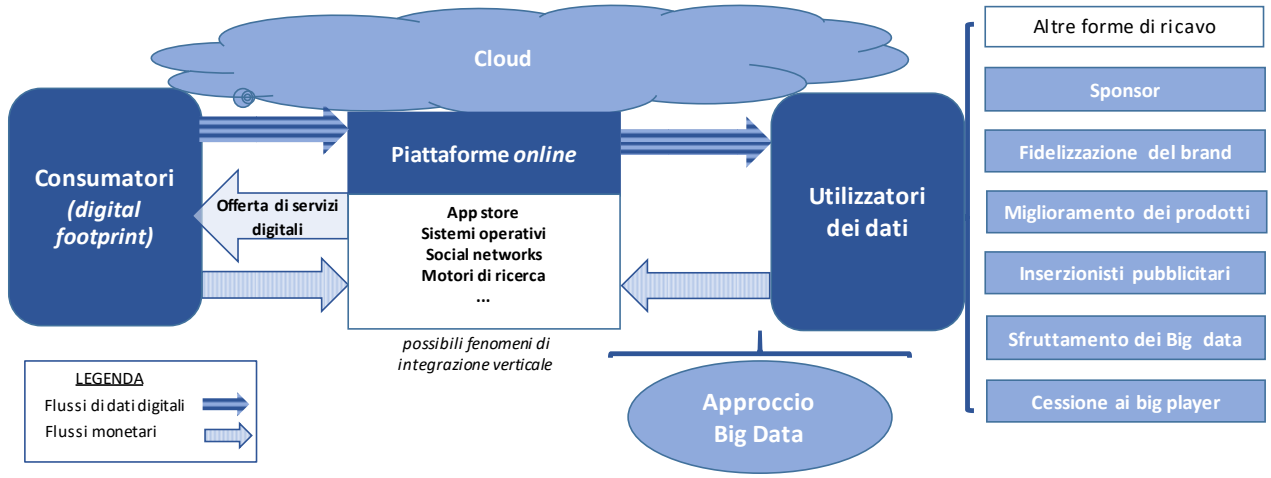


Figura 1.9 – Rappresentazione sintetica del mercato a due versanti applicato ai dati digitali
Fonte: Autorità

In terzo luogo, per quanto concerne il **versante degli utilizzatori di dati**, vale la pena osservare come questo ambito sia spesso caratterizzato dalla **presenza di più fasi (e a volte intermediari), e, al contempo, dall'integrazione verticale di alcune grandi piattaforme in tutti gli stadi**.

In questi contesti di mercato, operano inoltre sia le classiche **esternalità di rete (cd. "dirette")**, giacché per i singoli consumatori il valore del servizio offerto dalla piattaforma (es. un servizio di messaggeria, un *social network*,…) spesso aumenta al crescere della "rete" di consumatori che ne fa uso, sia quelle tipiche

di un mercato a più versanti, vale a dire le **esternalità di rete incrociate**, che si verificano quando le decisioni prese dagli attori appartenenti ad un lato del mercato producono effetti sugli agenti che fanno parte degli altri versanti, e la cui intensità incide in maniera determinante sulla struttura dei prezzi (per un maggiore dettaglio si rinvia al paragrafo 2.7). **Tali caratteristiche (anche dette "rendimenti di scala dal lato della domanda"), a parità di altre condizioni, fanno convergere naturalmente l'ecosistema verso esiti di mercato particolarmente concentrati.**

Ulteriori caratteristiche che riguardano la struttura dell'intero sistema dei *big data*, e che di riflesso si ripercuotono sugli assetti competitivi dei vari segmenti, sono diretta conseguenza delle *3 V* (v. paragrafo 1.1). Si tratta, in particolare, della **presenza di economie di scala (dal lato dell'offerta), come conseguenza della crescita del volume dei dati e della loro capacità di produrre rendimenti di scala crescenti, e di economie di scopo (o varietà), a seguito della crescente possibilità di combinare una varietà sempre maggiore di dati.** Questa struttura dei costi (con costi medi decrescenti e costi marginali bassi, quasi nulli), e la connessa distribuzione delle imprese (in letteratura nota come "*firm size distribution*"), è resa ancora più asimmetrica (ossia *skewed*) dalla presenza di ingenti costi fissi e affondati relativi alle attività di R&S (v. *infra*).

Ciò non può che ripercuotersi nella presenza di elevate **barriere all'entrata** (o più spesso **allo sviluppo**) o **all'accesso ai *big data***. In un'epoca in cui i processi decisionali sempre più si fondano sull'utilizzo di dati, l'analisi relativa all'accesso a questo rilevante *input* appare cruciale per le dinamiche dei mercati direttamente (es. pubblicità online) e indirettamente (es. informazione in rete) interessati. Le imprese che hanno una maggiore possibilità di raccogliere dati digitali, e/o quelle che riescono ad aggregare in maniera efficiente *dataset* eterogenei, e/o che possiedono le competenze e gli strumenti di *data analytics*, godono di un considerevole vantaggio concorrenziale. Le barriere all'entrata o allo sviluppo, infatti, individuano quelle circostanze che rendono difficile l'ingresso o l'espansione delle imprese in specifici mercati, garantendo a quelle che già vi operano un maggior potere di mercato.[33] **Barriere all'entrata e allo sviluppo sono riscontrabili in tutte le fasi della catena del valore e possono avere natura tecnologica, legale e/o strategica e si possono presentare anche contemporaneamente, rafforzandosi reciprocamente.**

In particolare, il fenomeno delle barriere all'entrata o all'espansione è dirimente nel primo stadio della catena del valore, ossia quello della raccolta dei dati, in conseguenza della forte dipendenza dallo stesso dei successivi stadi e quindi degli effetti di *spillover* che la creazione di una barriera nello stadio iniziale genera in quelli successivi. [34]

Alla luce di quanto descritto fino ad ora, appare evidente che la questione relativa all'accesso ai dati (modalità, rapporti operatori-consumatori, struttura di mercato, ecc.) diviene di primaria importanza al fine di creare contesti socialmente efficienti. Peraltro, in tali contesti si scontrano anche interessi collettivi differenti, quali privacy, concorrenza, pluralismo informativo.

Nella restante parte di questo Capitolo, si approfondiranno gli aspetti relativi al funzionamento dell'ecosistema, facendo riferimento alla componente relativa alla struttura tecnologica, di costo e

[33] Per un approfondimento sul ruolo delle barriere all'accesso presenti nel mercato dei *Big data* si veda, tra l'altro, il lavoro di RUBINFELD D.L., GAL M.S., (2017), *Access barriers to Big data*, 59 Arizona Law Review 339.

[34] A titolo di esempio, barriere all'entrata si possono verificare nella fase di raccolta dati laddove non per tutti gli operatori è possibile replicare la raccolta in parallelo dei dati come accade, ad esempio, per le informazioni di natura esclusiva che vengono raccolte dai *social network*. Per quanto riguarda la fase della conservazione dei dati, se è vero che si sono ridotti i costi delle tecnologie che consentono la loro archiviazione, è altrettanto evidente che l'esplosione del *cloud*, con i suoi connessi effetti di *lock-in*, rende difficile l'ingresso di nuove imprese nello specifico segmento di mercato. Infine, non di minore intensità sono gli effetti prodotti dall'esistenza di barriere all'accesso nelle fasi dell'analisi e dell'uso dei *big data*; come più volte ribadito, infatti, tali attività richiedono figure professionali di altissimo profilo e tecnologie (anche algoritmiche) assai complesse. Relativamente all'uso dei dati, una barriera di tipo legale, che merita menzione, è relativa alla corretta attribuzione dei diritti di proprietà.

concorrenziale dei versanti successivi alla fase iniziale di acquisizione del dato (e quindi alla relazione utente-operatore). Nei Capitoli successivi, si approfondirà invece questo primo stadio, con riferimento agli aspetti economici (v. paragrafo 2.4), di assegnazione efficiente e trasparente dei diritti di proprietà (v. paragrafo 2.7), nonché alle conseguenze sull'intero contesto politico-sociale (v. Capitolo 3).

1.5. Le analisi di specifici segmenti

Il passo successivo dell'analisi si pone l'obiettivo di studiare alcuni specifici segmenti dell'ecosistema dei *big data* allo scopo di mostrare come le caratteristiche strutturali poc'anzi descritte si concretizzano, dal lato produttivo e tecnologico, in scenari e assetti di mercato.

1.5.1. Il primo livello: i sistemi operativi

Un primo interessante ambito, analizzato sia in letteratura sia nella casistica regolamentare e antitrust, riguarda **i sistemi operativi (SO). I sistemi operativi rappresentano un canale (*gatekeeper*) privilegiato per la raccolta di dati, palesandosi, di fatto, come una vera e propria barriera all'entrata di tipo tecnologico.**

Infatti, i sistemi operativi sono software che controllano le funzioni di base di un *device* e consentono all'utente di utilizzare il *device* stesso e le applicazioni software installate. Sono sviluppati da produttori specializzati (come nel caso di Microsoft), ovvero da società che producono anche l'hardware (v. *Apple*). I sistemi operativi possono controllare la funzionalità di qualsiasi tipo di apparecchio di navigazione web, dai tradizionali pc, ai nuovi dispositivi mobili (smartphone, tablet, …).

In questo senso, i sistemi operativi si pongono in cima nella gerarchia dei livelli attraverso cui l'individuo naviga in rete (**Figura 1.10**), con ciò caratterizzandosi come l'ambito più diretto e principale per l'acquisizione della *digital footprint* (v. Capitolo 2).

In generale, il settore dei sistemi operativi è dominato dalla presenza di pochi operatori, peraltro verticalmente integrati, ed è segmentato in due (o più) mercati merceologici distinti: i sistemi operativi per pc e quelli per *device* mobili. Tali mercati presentano diverse strutture, anche in considerazione del fatto che attraversano una diversa fase del ciclo di vita del prodotto.

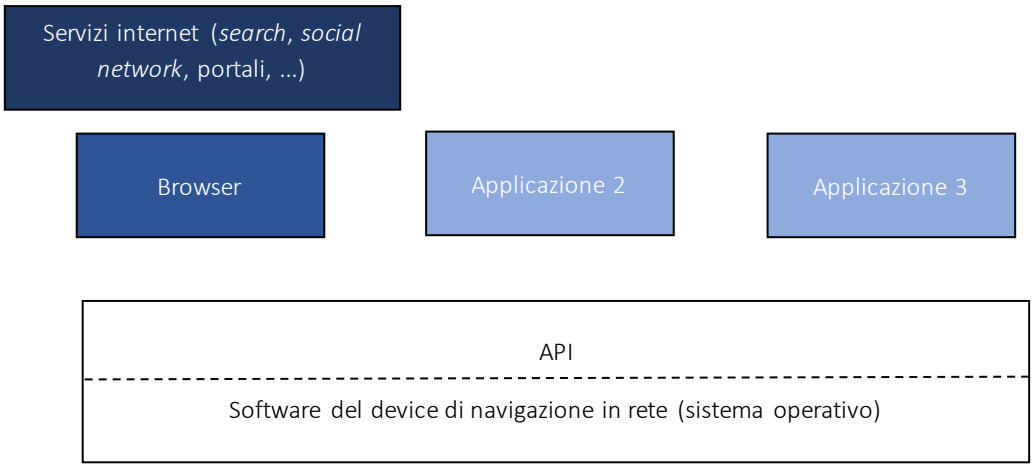


Figura 1.10 – Stadi nell'accesso ai dati individuali
Fonte: Autorità

Di particolare interesse per la presente analisi, per una molteplicità di motivazioni che saranno illustrate nel Paragrafo 2.5 e seguenti (a cui si rimanda), è il mercato dei **sistemi operativi per *device* mobili**. In questo ambito, le dinamiche di diffusione dei principali sistemi operativi mobili nel mondo (**Figura 1.11**) mostrano chiaramente la prevalenza di (soli) due sistemi operativi, quello che va sotto il nome di *Android*, prodotto da *Google*, e il sistema operativo *iOS* prodotto da *Apple*. Il primo è un sistema operativo gratuito e *open source*, il secondo è il sistema operativo sviluppato da *Apple*, ed è utilizzabile solo attraverso i *device* prodotti e distribuiti dalla società di Cupertino (*iPhone*, *iPod touch* e *iPad*).

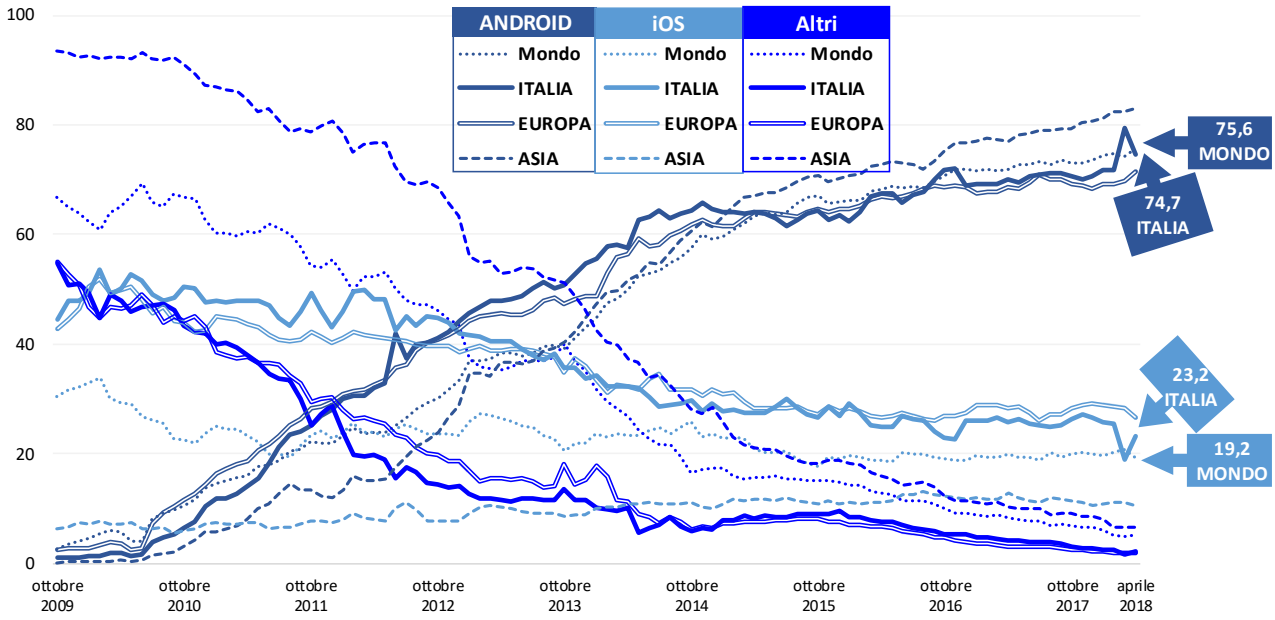


Figura 1.11 – Diffusione dei sistemi operativi per dispositivi mobili nel mondo (ottobre 2009 – aprile 2018)
Fonte: Elaborazioni AGCOM su dati mensili *StatCounter.com*

Google e *Apple*, quindi, sono le aziende *leader* tra i sistemi operativi per *device* mobili; ad aprile 2018, la quota cumulata a livello mondiale arriva a sfiorare il 95% (98% in Italia), in aumento rispetto al 90% toccato ad ottobre 2016. Considerando la dinamica nel tempo, la quota di mercato del sistema operativo *Android* mostra un *trend* di forte crescita in tutte le aree geografiche considerate, fino a rappresentare il 75,6% dei sistemi operativi nel mondo (74,7% in Italia). La dinamica del sistema operativo *iOS*, invece, mostra un *trend* in leggera flessione che ha portato la quota di mercato ad assestarsi intorno al 20%. In generale, il mercato dei sistemi operativi dei *device* mobili, così come quello delle postazioni *desktop*, è caratterizzato da un elevatissimo livello di concentrazione, che sfocia in un assetto duopolistico su tutti i mercati mondiali; tale concentrazione è principalmente dovuta, come detto, all'esplicarsi degli effetti di rete[35].

In questo quadro, tutti gli altri sistemi operativi mostrano un andamento in declino che li ha portati quasi alla scomparsa; ciò a partire dall'introduzione sul mercato, nel 2008, del primo smartphone di *Apple*. Fino al 2011, la loro complessiva penetrazione era ancora superiore a quella dei due sistemi operativi

[35] In tal senso, "*Early on, [Microsoft] recognized that consumers would benefit greatly if a wide range of hardware and software products could interoperate with one another. Among other things, (i) the products would be more useful if information could be exchanged among them, and (ii) development costs would fall and a broader array of products would become available if they could be developed for larger customer segments without the need to rewrite software to target narrow platforms. As more products became available and more information could be exchanged, more consumers would be attracted to the platform, which would in turn attract more investment in product development for the platform. Economists call this a "network effect," but at the time we called it the "positive feedback loop"*." Cfr. testimonianza diretta di Bill Gates, Civil Action No. 98-1233 (CKK), paragrafo 25.

attualmente più diffusi, mentre da quell'anno in poi si è potuto assistere a una loro graduale e rapida riduzione, fino a rappresentare, ad aprile 2018, appena il 5% dei sistemi operativi in circolazione.[36]

Sulla diffusione dei sistemi operativi incidono, inoltre, in maniera determinante, da una parte, il livello di integrazione verticale con i produttori di *device*, in particolare i produttori di smartphone (si veda il caso *Apple*), e, dall'altra, il livello di apertura dei software medesimi. Il sistema operativo *Android* si presenta con un grado di apertura superiore a quello di *iOS* ed è quindi utilizzabile per il funzionamento di numerosi dispositivi mobili offerti da diversi produttori, mentre il sistema operativo *iOS* viene impiegato solo dai prodotti di *Apple*.

In ogni caso, il primo livello nella catena dell'ecosistema *big data*, almeno nella parte riguardante i dati direttamente associati agli individui, è caratterizzato da un'elevata e crescente concentrazione dei mercati.

1.5.2. Il secondo livello: i motori di ricerca e i social network

Un secondo livello di acquisizione dei dati riguarda quello relativo alla navigazione in rete che può avvenire attraverso specifici software (i *browser*) o direttamente attraverso APP. In questo contesto, motori di ricerca e *social network* rappresentano piattaforme che concentrano su di esse una rilevanza assoluta in termini di audience (*reach* e tempo speso dai consumatori),[37] di rilevanza sociale, e di preminenza dal punto di vista del pluralismo informativo (v. Capitolo 3).

Su tali specifici ambiti di mercato l'Autorità già da tempo ha intrapreso studi ed analisi, data la rilevanza che tali strumenti assumono oggi nelle modalità attraverso cui si forma l'opinione pubblica.[38]

Per quanto riguarda il ***search***, questo è uno dei primi servizi online ad essere stato offerto una volta che il web è diventato un sistema aperto ai contenuti e ai servizi di privati.[39] I motori di ricerca risolvono problemi transazionali, sia dal lato della domanda (la ricerca dell'utente di informazioni), sia dal lato dell'offerta (la necessità dei soggetti che offrono servizi e prodotti di essere noti agli utenti), svolgendo pertanto il ruolo di piattaforma. **In particolare, il *search* è caratterizzato dall'esistenza di esternalità di rete incrociate (o effetti di feedback tra i due versanti del mercato), che contribuiscono a determinare un esito di mercato particolarmente concentrato.**

L'andamento storico delle quote di mercato (**Figura 1.12**) permette di osservare un'evoluzione tipica dei mercati in cui prevalgono le esternalità di rete: una prima fase in cui operano più soggetti, seguita da un fenomeno concentrativo in cui si afferma una piattaforma dominante (con quote superiori all'80%).

[36] Nella voce "Altri" sono compresi numerosi sistemi operativi tra cui *BlackBerry OS*, *Series 40*, *Windows Phone*, *Samsung*, *SymbianOS*. Alcuni di questi sistemi operativi sono scomparsi.
[37] V. Osservatorio sulle comunicazioni, n.1/2018, (slide 2.4); https://www.agcom.it/documents/10179/10293149/Studio-Ricerca+16-04-2018/1c205715-a147-43fa-a9a1-f778690fe65b?version=1.3.
[38] *Indagine conoscitiva sui servizi internet e sulla pubblicità online*, conclusa con delibera n. 19/14/CONS https://www.agcom.it/documents/10179/1/document/9376a211-ebb2-4df6-83ea-282f731faaf2 .
[39] Per la storia dei motori di ricerca v. www.searchenginehistory.com, www.wordstream.com

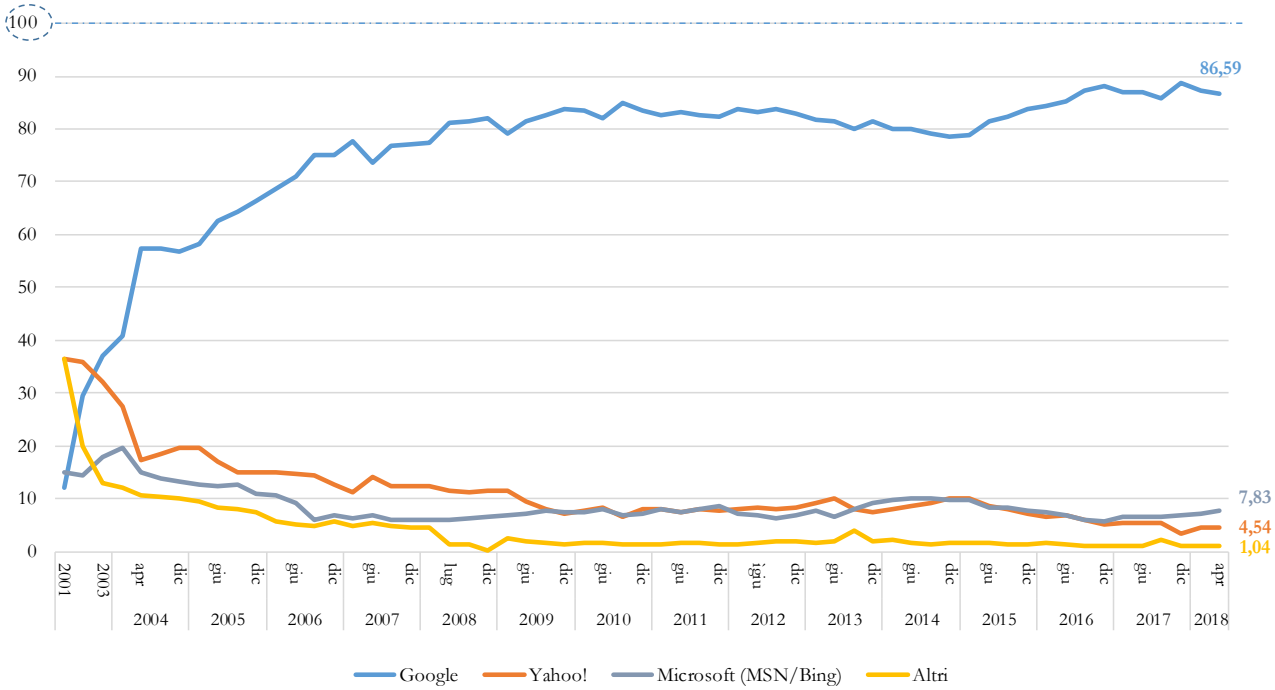


Figura 1.12 – Evoluzione storica delle quote di mercato dei motori di ricerca nel mondo (%)
Fonte: elaborazione Agcom su dati *SEW/WebSideStory, NetApplications, NetMarketShare e StatCounter*

L'assetto di mercato del *search* si è, quindi, sviluppato verso una struttura particolarmente concentrata, in cui, allo stato attuale, un operatore, *Google*, detiene, oramai da quasi un quindicennio e quasi ovunque nel mondo, quote di mercato superiori all'80%[40].

Il secondo caso considerato è quello relativo ai ***social network***; come noto, questi costituiscono piattaforme online che consentono agli utenti di costruire un profilo pubblico o semi-pubblico all'interno di un sistema predefinito, creando una propria rete di contatti (tra gli utenti collegati e iscritti alla medesima piattaforma), nonché di visualizzare e scorrere le liste di utenti presenti negli altri profili. La natura e la definizione delle relazioni nell'ambito della rete sociale di utenti possono variare da un sistema a l'altro.

I ***social network*** si distinguono dagli altri servizi (quali messaggerie, chat, blog) non tanto per la possibilità di interagire e incontrare nuove persone, quanto piuttosto per avere una rete sociale di contatti visibile.

Numerose sono le caratteristiche che consentono di distinguere i vari *social network* fra cui, a titolo non esaustivo, si possono menzionare il diverso grado di visibilità pubblica del profilo, e dei relativi contatti, gli strumenti disponibili per interagire, la base di utenti (il *target*) cui si rivolge il servizio. Inoltre, un'ulteriore distinzione avviene in relazione alle funzionalità offerte ai propri utenti, tra cui la possibilità di lasciare commenti, messaggi, suggerimenti, nonché di esprimere la propria reazione, pur utilizzando modalità e denominazioni diverse. Inoltre, fra le funzionalità tecnologiche più diffuse si ricorda la condivisione di foto, video e contenuti multimediali, la creazione di blog e forum di discussione, *l'instant messaging* e le chat. **Tutte attività che comportano la generazione di dati digitali, che vengono raccolti ed elaborati dai *social network*** (in tal senso, v. Capitolo 3).

Relativamente alle caratteristiche di mercato, il modello di *business* scelto per i *social network* è caratterizzato, in modo del tutto simile agli altri servizi web di tipo orizzontale, dalla

[40] Per una disamina concorrenziale del *search*, e dei *social network*, così come di altri ambiti di mercato (sistemi operativi, *browser*, portali,…) si rimanda al Capitolo 3 della citata *Indagine conoscitiva sui servizi internet e sulla pubblicità online.*

valorizzazione dei contatti in termini pubblicitari, a fronte di un servizio completamente gratuito (o quasi) per gli utenti.

Nonostante esistano esempi di (parziale) valorizzazione sul versante degli utenti finali (es. *LinkedIn*) e/o degli sviluppatori di programmi e applicazioni (es. *Facebook*), tuttavia, per tutti i *social network* la componente pubblicitaria risulta ancora predominante. Al riguardo, si è osservata **una dinamica evolutiva molto simile a quella del *search*** (

Figura 1.13), nella quale l'acquisizione della massa critica di utenti necessaria allo sfruttamento degli effetti di rete e a innescare fenomeni postivi di retroazione ha rappresentato l'elemento indispensabile per l'affermazione della piattaforma, nonché il presupposto al raggiungimento e superamento del *break even* prevalentemente attraverso forme (innovative) di pubblicità. Grazie all'operare degli effetti di rete, nel giro di pochi anni, ossia dal lancio del sito (avvenuta, come detto, nel 2006) ad oggi, *Facebook* ha rapidamente raggiunto la leadership mondiale con una quota di mercato che è stata stabilmente superiore all'80% per tre anni (giugno 2014 – ottobre 2017), per poi diminuire leggermente ed assestarsi, attualmente, su una quota superiore al 70%.

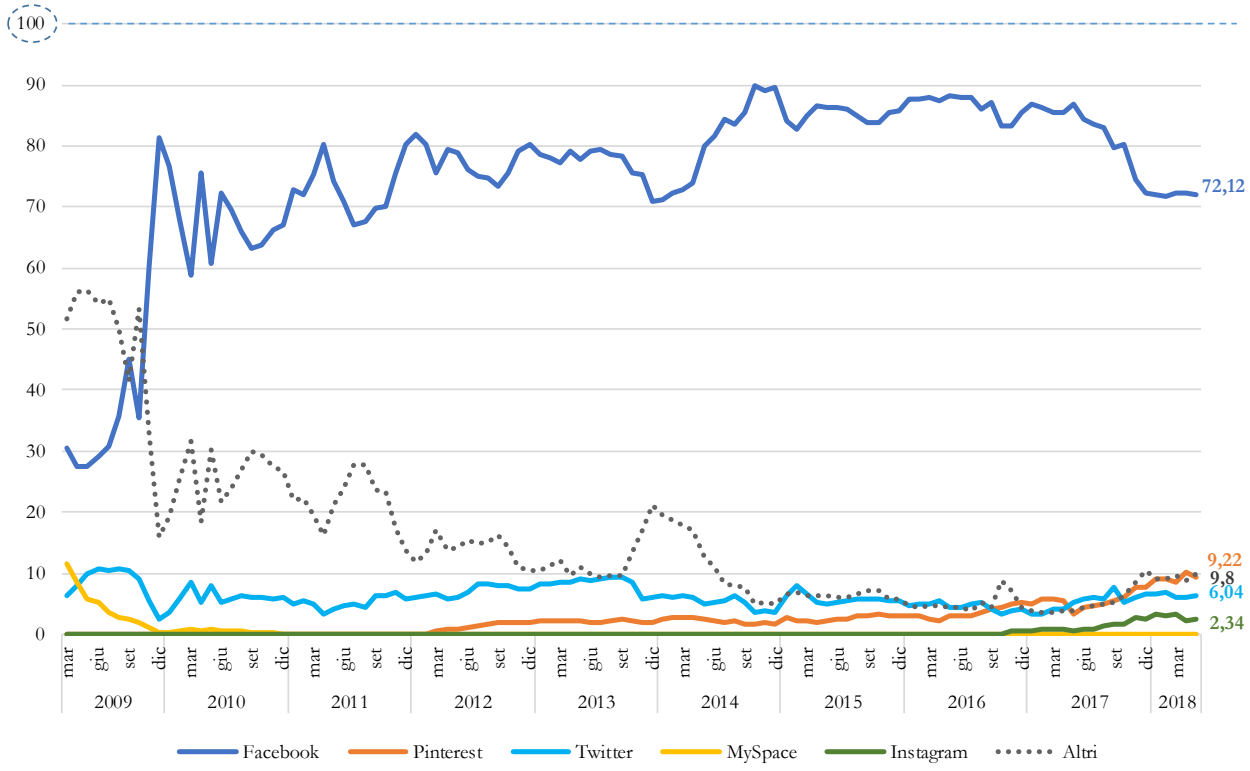


Figura 1.13 – Evoluzione storica delle quote di mercato dei *social network* in Europa (%)
Fonte: elaborazione Agcom su dati *StatCounter*

Questo processo è stato facilitato dall'interconnessione con altri siti internet in grado di "richiamare" traffico dati e aumentare il coinvolgimento degli utenti. L'interconnessione tra servizi internet rappresenta, infatti, un elemento chiave per l'affermazione di processi evolutivi di scala, nonché per la riduzione delle barriere all'entrata, nel momento in cui, qualche operatore si sia già affermato.

Inoltre, il declino improvviso di alcuni *social network* – si pensi, ad esempio, a *Friendster*, con riferimento al quale è stata sufficiente la perdita di alcuni (gruppi di) utenti per comportare un abbandono di massa della piattaforma – conferma l'assoluta rilevanza per tale tipologia di servizio web degli effetti di rete diretti.

1.5.3. Il terzo livello: i data center ("La capacità produttiva")

Un ultimo caso utile a mostrare come le caratteristiche della struttura di mercato dei *big data* possano portare a equilibri di mercato che presentano alti livelli di concentrazione è quello relativo alla creazione delle infrastrutture materiali che consentono l'attività di acquisizione e immagazzinamento dei dati; si tratta del segmento di mercato legato ai ***data center***.

Un *data center* rappresenta un'infrastruttura fisica (un edificio o un immobile), che ospita le apparecchiature di elaborazione dei dati (ovvero i *server* e l'infrastruttura necessaria per il loro funzionamento) di una o più società o organizzazioni (co-locazione).[41] Il *data center*, quindi, deve essere costituito da almeno una stanza separata con alimentazione elettrica indipendente e relativa climatizzazione. Questa definizione consente immediatamente di mettere in evidenza che si tratta di un segmento di mercato in cui operano piccole, medie e grandi imprese. In molti casi, la stessa attività operativa delle imprese necessita di *data center*, le cui dimensioni, quindi, dipendono dalle esigenze commerciali della società.[42]

I *data center* rappresentano la principale infrastruttura dell'economia *data-driven*. L'offerta di un qualsiasi servizio online si appoggia su *server* ubicati in *data center*. I *data center* più imponenti sono strutture complesse con numerose strumentazioni meccaniche, elettriche e di comunicazione. Al fine di un efficiente funzionamento, i *data center* necessitano di energia elettrica per l'alimentazione dei *server*, e acqua per il loro raffreddamento. Inoltre, per trasportare i dati da e verso i *data center* sono necessarie infrastrutture di comunicazione (di solito collegamenti in fibra ottica).

Gli investimenti in *data center*, di conseguenza, sono costantemente in aumento, spinti sia dalla necessità delle imprese private nella costruzione di propri centri per la raccolta e la elaborazione dei dati, sia dalla intensa diffusione dei servizi di ***cloud computing***, ossia l'offerta esternalizzata di servizi di immagazzinamento e calcolo (server, risorse di archiviazione, database, sofware, analisi, ecc.) tramite web.

È bene sottolineare che molti operatori che forniscono servizi digitali costruiscono *data center* principalmente per poter gestire i dati che acquisiscono direttamente dai propri utenti; è questo il caso di *Google*, che comunque compare anche tra i principali *player* del *cloud* e di *Facebook*, che al momento non contempla (se non indirettamente), nella propria offerta, la vendita di servizi di *cloud* ma che ovviamente dispone di propri *data center*.

---

[41] Secondo uno studio prodotto dalla *Microsoft* nel 2009, i costi complessivi di un di un *data center* risultano distribuiti in percentuale come segue; 45% per i server e altre componenti (CPU, memorie, sistemi di *storage*), un 25% dei costi sono da imputare alla creazione di infrastrutture per la trasmissione di energia per il funzionamento ed il raffreddamento dei server, un 15% sono relativi all'suo di corrente elettrica ed il restante 15% sono relativi alla creazione della rete (*network*) tramite il quale si collega alla rete un *data center*. Cfr. GREENBERG A., HAMILTON J., MALTZ D.A., PARVEEN P., (2009), *The Cost of a Cloud: Research Problems in Data Center Networks*, Microsoft Research, Redmond USA.

[42] Ad esempio, un'impresa come Eni nel 2013 ha inaugurato il suo data center (a Ferrera Erbognone, nel cuore della Pianura Padana) nel quale vi sono gli strumenti fondamentali (il supercalcolatore HPC4 capace di svolgere 22,4 milioni di miliardi di operazioni matematiche in un secondo) per le scoperte esplorative di Eni in tutto il mondo e utili, per esempio, a elaborare immagini in 3D del sottosuolo. In totale oltre 7.000 sistemi, con più di 60.000 core CPU. https://www.eni.com/it_IT/innovazione/piattaforme-tecnologiche/aumento-recupero-idrocarburi/hpc.page.
Un esempio che riguarda un'impresa di piccole dimensioni, ma che comunque mostra quanto importante siano le specifiche di un *data center*, riguarda la società *CriptoMining* (http://criptomining.online/) una startup innovativa che produce criptovalute su scala industriale attraverso un processo informatico noto come *mining*. Il *data center* Situata si trova nei sotterranei di un edificio nel centro di Milano e impiegherà a regime 250 macchine h24 (al momento risultano operative 12 macchine). La struttura è ubicata tre piani sotto terra, per garantire basse temperature che consentono il raffreddamento delle macchine, sorvegliata 24 ore su 24.

Secondo *lifelinedatacenters.com*, i *data center* sono circa 8,6 milioni nel 2017, seppure in diminuzione per effetto della diffusione sempre maggiore dei servizi di *cloud*; a tale diminuzione fa da contraltare la crescita dei metri quadri necessari per ciascun singolo impianto, vale a dire un aumento della dimensione media[43], a cui si associano forti economie di scala.[44]

Ad esempio, dal 2007 al 2017, *Google* ha investito circa 3,2 miliardi di euro per costruire i quattro *data center* che operano attualmente in Europa, per un valore di circa 300 milioni di euro l'anno;[45] ciò implica l'esistenza di crescenti ed elevati costi di ingresso per i soggetti che vogliono entrare in questi ambiti di mercato.

Il fenomeno dei *data center* non è recente, dal momento che tutte le imprese da sempre hanno avuto bisogno di spazi fisici in cui archiviare le informazioni. Risulta tuttavia evidente come, sotto la pressione esercitata dai *big data*, l'esigenza di utilizzare sempre più infrastrutture di acquisizione e gestione dei dati (anche in *cloud computing*, **Figura 1.14**) abbia di fatto costituito una nuova e crescente spinta verso questo tipo di infrastrutturazione.

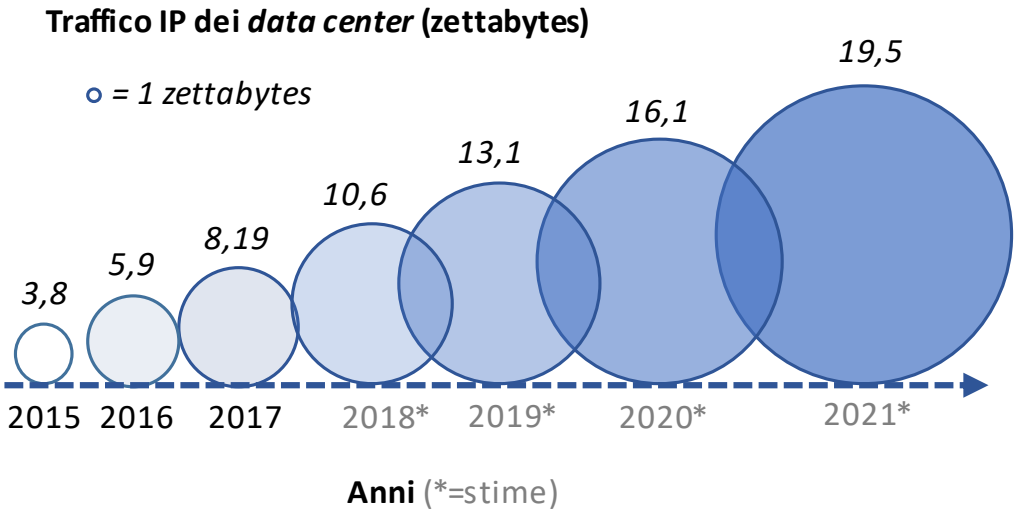


Figura 1.14 – Crescita del traffico IP per servizi di *cloud*
Fonte: Elaborazioni AGCOM su dati *Cisco*

L'enorme diffusione della *big data economy*, e di conseguenza del *cloud* (Figura 1.14), ha prodotto delle modifiche sostanziali in un segmento precedentemente dominato dai *data center* proprietari. Lo sviluppo del segmento, inoltre, risulta fortemente **influenzato dalle stesse caratteristiche dei *big data*** delineate in precedenza**.**

La dimensione, la varietà e la velocità dei dati hanno spinto l'ecosistema verso una struttura di costi dominata dalla presenza di economie di scala, e quindi da un assetto concorrenziale contraddistinto da crescenti livelli di concentrazione nonché da barriere all'ingresso sempre maggiori.

[43] https://lifelinedatacenters.com/data-center/emerging-data-center-trends/
[44] cfr. GREENBERG A., HAMILTON J., MALTZ D.A., PARVEEN P., (2009), *The Cost of a Cloud: Research Problems in Data Center Networks*, Microsoft Research, Redmond USA.
[45] Fonte: *European data centres* a cura di Copenahgen Economics, 2018. Si tratta dei *data center* localizzati a Dublino (Irlanda), St. Ghislain (Belgio), Eemshaven (Olanda) e Hamina (Finlandia).

Secondo la società di consulenza *Gartner,* il mercato mondiale dei servizi di (*public*) *cloud*[46] è previsto quest'anno in crescita del 21,4% per un ammontare totale di ricavi pari a 186 miliardi di dollari. In questo contesto, i 10 provider principali hanno aumentato sensibilmente la propria quota di mercato, arrivando a detenere congiuntamente il 70%.[47]

Analogamente, secondo uno studio fornito dalla società *Synergy Research Group* (**Figura 1.15**), il mercato complessivo (*public* e *private cloud)* è caratterizzato, con una quota in crescita e pari al 34%, dalla *leadership* di Amazon, che offre il servizio *Amazon Web Services*, ed è stato il primo operatore, nel 2006, a lanciare un servizio di *cloud* su larga scala.

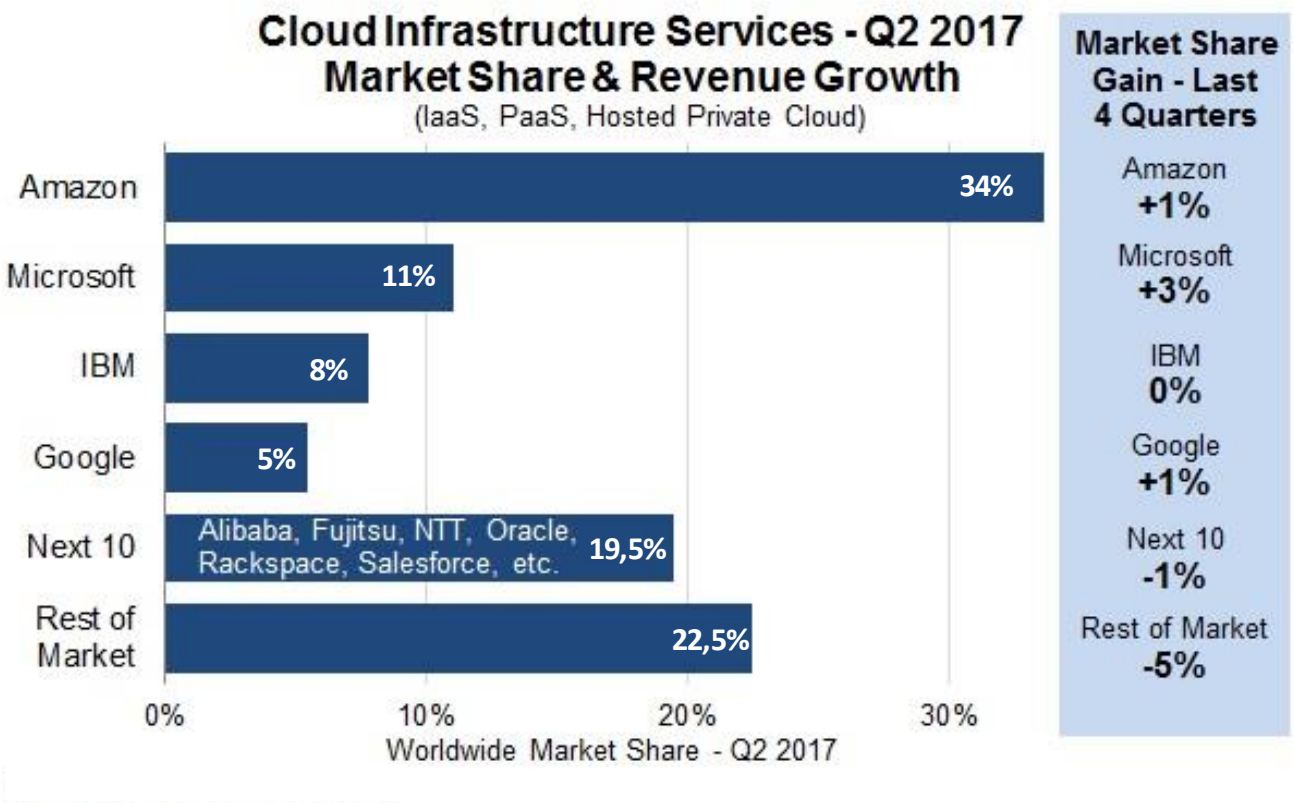


Figura 1.15 – Quote di mercato nei servizi di *cloud* (2° trimestre 2017)
Fonte: https://www.srgresearch.com/articles/leading-cloud-providers-continue-run-away-market

La struttura infrastrutturale segue poi la dimensione globale che hanno assunto i maggiori provider. L'esistenza di economie di scala (e di scopo), nonché l'esigenza di duplicazione dei dati (v. *supra*) e di

[46] La differenza principale tra *private* e *public cloud* risiede nel fatto che, optando per la seconda soluzione, i dati sono immagazzinati nei *data center* di chi offre un servizio esternalizzato che, quindi, risulta responsabile della loro gestione e manutenzione. Se tra le grandi aziende l'utilizzo del *public cloud* è ancora poco diffuso, giacché si preferisce mantenere il controllo diretto sui dati, per molte piccole imprese e soprattutto per i singoli utenti (che rappresentano comunque una fonte primaria di dati digitali) l'utilizzo di questi ambienti di archiviazione è divenuto indispensabile.
Il *public cloud* individua una particolare architettura che garantisce a ciascun utente uno spazio di archiviazione personale (*cloud storage*) che presenta la caratteristica di essere accessibile in qualsiasi luogo e con qualsiasi *device* tramite una connessione ad internet, a cui possono essere affiancati anche servizi di elaborazione dei dati.
Tramite il *cloud storage*, è possibile sincronizzare tutti i propri file in un unico posto (la nuvola), con il conseguente vantaggio di poterli ri-scaricare, modificare, cancellare, aggiornare, senza avere quindi più il bisogno di portare con sé le così dette memorie esterne (*hard disk* esterni, *pen drive* USB, ecc.).
Con l'avvento dell'approccio *cloud* si è quindi assistito a un radicale cambiamento nella maniera in cui vengono trattati i dati sia dai consumatori che dagli operatori.
[47] https://www.gartner.com/newsroom/id/3871416.

collocazione di centri di calcolo sempre più vicini ai consumatori hanno infatti spinto gli operatori a creare delle vere e proprie reti globali (si veda ad esempio quella di *Google*, **Figura 1.16**).

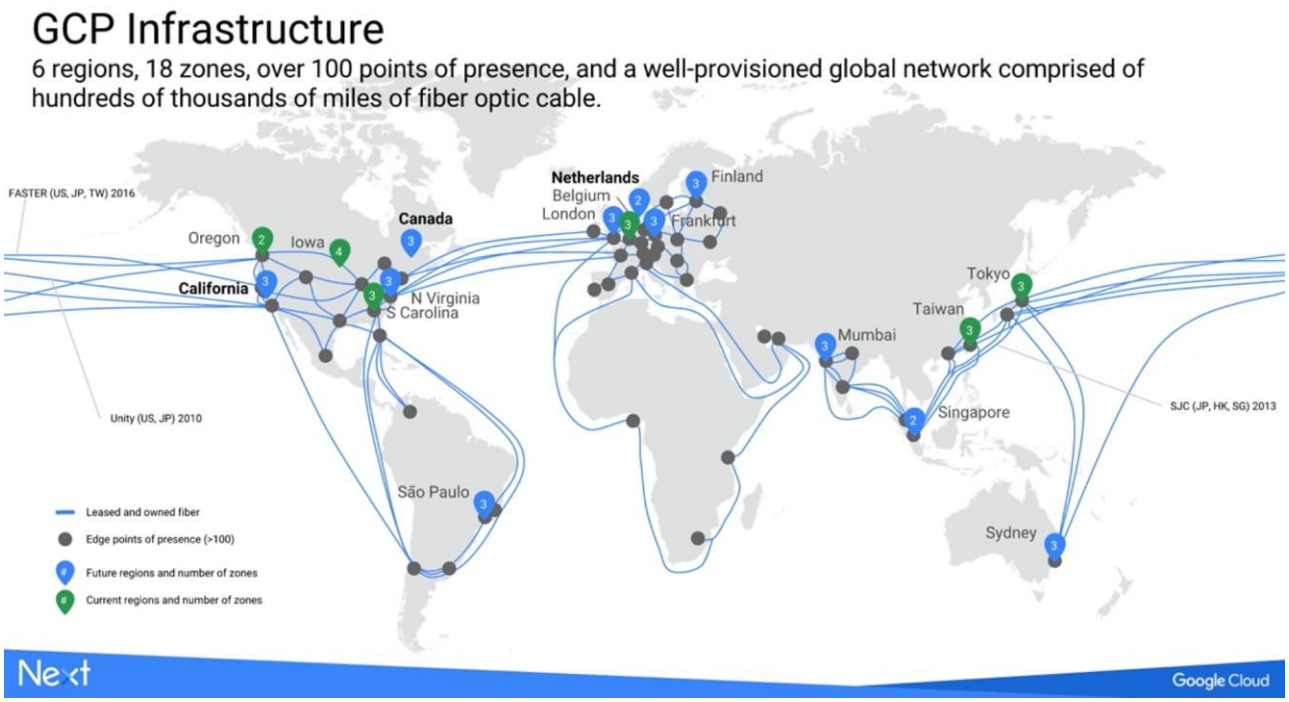
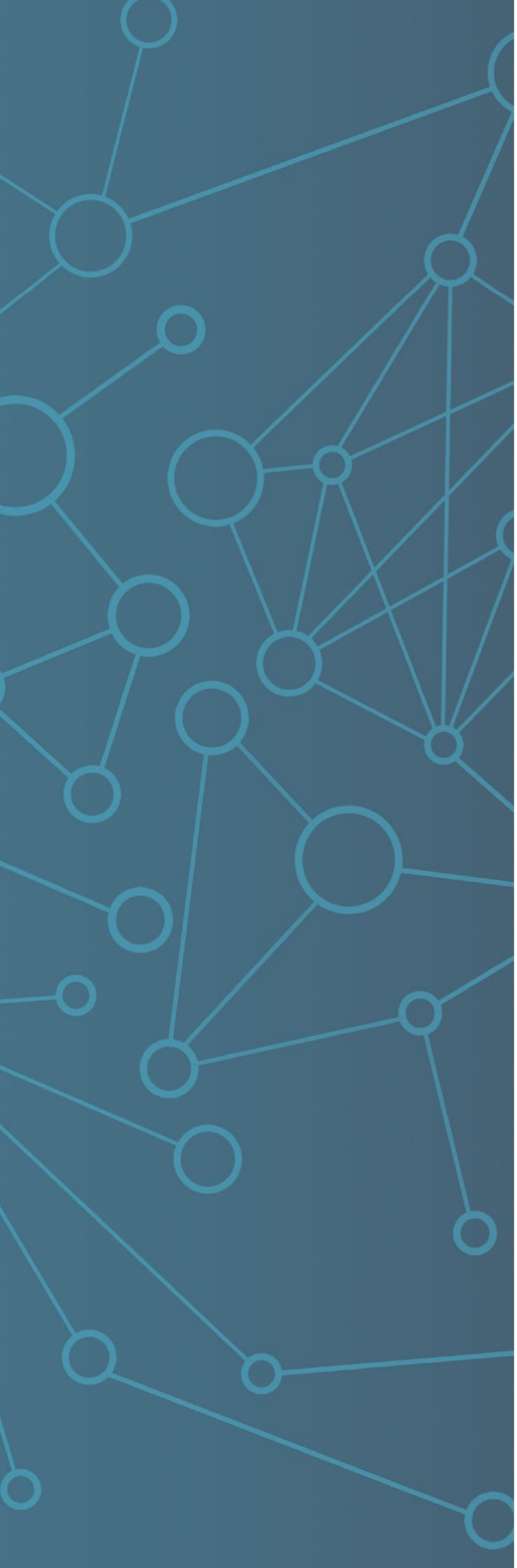


Figura 1.16 – Infrastruttura di *Google* per la fornitura di servizi cloud - *Google Cloud Platform* (GCP) – (2017)
Fonte: *Google Cloud Next* 2017, 8 - 10 marzo 2017 San Francisco - https://www.youtube.com/watch?v=vX92qwNtkFo&t=1362s

Relativamente ai costi, essi appaiono di difficile determinazione. Secondo uno studio del 2014, le dimensioni medie dei *data center* di *Google* sono comprese tra i 15.000 e i 18.500 m^2, seppure alcuni raggiungono i 90.000 m^2 come quello di *Pryor Creek* in Oklahoma (USA). La variabilità della dimensione, inoltre, rappresenta una specifica strategia adottata dalle compagnie, che strutturano i propri *data center* a seconda delle necessità e della specifica area geografica.[48] Non a caso, nel 2003, *Google* ha ottenuto anche un brevetto per la realizzazione di *data center* modulari (costruiti in nei classici *container* utilizzati per il trasporto delle merci) che consentono una maggiore elasticità nella realizzazione delle infrastrutture. È interessante anche notare che la potenza energetica, in megawatt (MW), necessaria per il funzionamento dei *data center* di *Google* è stata stimata, in un anno, pari allo 0,01% della potenza energetica mondiale.[49]

In conclusione, l'avvento dei *big data* ha determinato l'acquisizione, l'immagazzinamento e l'analisi di un numero e di una varietà crescente di dati. Ciò sta avendo un forte impatto sulla sottostante struttura dei costi e quindi sui relativi assetti di mercato. Nonostante il passaggio da un regime di scarsità a uno di ridondanza di dati, si stanno delineando equilibri di mercato assai concentrati, che, poi, non possono che riverberarsi anche sui collegati ambiti di mercato a valle. In ragione di queste e di altre caratteristiche (l'esistenza in particolare di esternalità di rete), ambiti quali i sistemi operativi, i servizi web (quali *search* e *social network*) e la pubblicità online presentano caratteristiche di forte e crescente concentrazione.

[48] GHIASI A., BACA R., (2014), *Overview of largest Data Centers*, 802.3bs Task Force.
[49] GHIASI A., BACA R., (2014), *Overview of largest Data Centers*, 802.3bs Task Force.



2

L'INDIVIDUO COME FONTE DI DATI

2.1. I soggetti attivi

La complessità sottostante la catena del valore determina uno scenario di mercato dei *big data* molto variegato e articolato. **Se gli attori che partecipano al mercato possono essere individuati, anche se non sempre con facilità, molto più difficile risulta sciogliere l'intricato intreccio di interazioni che avvengono nel mondo dei *big data*.**

Per quanto finora descritto, nell'ecosistema dei *big data*, è possibile identificare, tra gli altri, i seguenti attori principali:[50]

g) **i soggetti generatori di dati** (o fornitori di dati);
h) i **fornitori della strumentazione tecnologica**, tipicamente sotto forma di piattaforme per la gestione dei dati;
i) gli **utenti**, cioè coloro che utilizzano i *big data* per creare valore aggiunto;
j) i ***data brokers***, cioè le organizzazioni che raccolgono dati da una varietà di fonti sia pubbliche, sia private, e li offrono, a pagamento, ad altre organizzazioni;
k) le **imprese e le organizzazioni di ricerca**, la cui attività diventa fondamentale per lo sviluppo di nuove tecnologie, di nuovi algoritmi attraverso cui esplorare i dati ed estrarne valore;
l) gli **enti pubblici**, sia in qualità di enti regolatori dei mercati, sia con riferimento alle attività della pubblica amministrazione volte a migliorare i prodotti e i servizi offerti alla cittadinanza e in grado di aumentare il benessere collettivo.

Tuttavia, l'ecosistema dei *big data* presenta un **grado di interconnessione tra i vari soggetti che vi partecipano tale da rendere difficile l'identificazione di singoli mercati ben definiti**; la complessità che ne deriva, di conseguenza, determina uno scenario in cui i vari segmenti del sistema, di cui la **Figura 1.8** offre una possibile rappresentazione, risultano spesso tra loro strettamente interrelati. Ciò determina un assetto di mercato in cui operano **(poche) grandi imprese multinazionali, caratterizzate da un elevato grado di integrazione verticale, diagonale e orizzontale in tutte (o quasi tutte) le fasi dell'ecosistema, accanto a una miriade di piccole imprese specializzate** che spesso, dopo il periodo di *start-up*, vengono acquisite da quelle più grandi.

Gli assetti competitivi negli specifici ambiti di mercato risultano strettamente dipendenti da alcune comuni caratteristiche strutturali; tale circostanza giustifica un approccio, in prima battuta, di portata più ampia, teso ad individuare una serie di caratteristiche strutturali dei *big data*, per poi approfondire gli effetti della loro concretizzazione in alcuni specifici ambiti.

Vale inoltre rilevare che, come si illustrerà in maggior dettaglio nel Capitolo (v. paragrafo 2.6 e 2.7), **l'ecosistema dei *big data* è caratterizzato dalla presenza di numerose forme di contrattazione incompleta, da mercati impliciti (ossia in cui la contrattazione del bene avviene in maniera spuria), nonché da ambiti di tipo nozionale (ossia caratterizzati da perfetta integrazione verticale e da una domanda di mercato meramente potenziale).**

Ciò di per sé è già fonte di **fallimenti di mercato che pregiudicano l'efficienza sociale, statica e dinamica, dell'intero ecosistema dei *big data***.

[50] CURRY E., (2016), *The Big data value chain: definitions, concepts and theoretical approaches*, in New Horizons for a Data-Driven Economy, Springer, Cham.

2.2. I dati digitali e l'individuo

Un ambito particolarmente rilevante nei *big data* è quello delle informazioni riguardanti i singoli individui e, conseguentemente, di come queste informazioni sono tutelate. Tradizionalmente, vi è distinzione tra "dati personali" e informazioni che non sono considerate tali. Questo tipo di approccio se appare abbastanza semplice da comprendere dal punto di vista teorico, non lo è nella pratica, dal momento che **risulta sempre più difficile stabilire, come conseguenza della complessità del fenomeno *big data*, tra tutte le informazioni raccolte su un individuo cosa rappresenta un dato personale, cosa no.**

Molti dei problemi pratici dell'approccio tradizionale derivano dalla difficoltà di stabilire a priori quali siano i dati che consentono di identificare un individuo, le sue abitudini, anche quelle più private. Ciò è insito nella natura stessa di *big data*, che è stata investigata nel Capitolo precedente.

Lo sviluppo tecnologico e dalle modalità attraverso cui gli individui consumano prodotti e servizi tipici di un'economia digitale rendono, inoltre, l'acquisizione, la conservazione e l'analisi di grandi quantità di dati, strutturati e non, un'attività a cui difficilmente può essere posto un freno. La diffusione dell'uso di internet tra la popolazione mondiale, infatti, ha prodotto effetti dirompenti all'approccio tradizionale alle informazioni individuali; internet è uno strumento oramai pervasivo che per molti risulta essenziale sia in ambito di lavoro, sia per gli aspetti legati alla vita quotidiana (divertimento, fitness, turismo, lettura, ecc.). Avere a disposizione una connessione "*always on*" è spesso considerato dai cittadini un elemento indispensabile della propria vita[51].

Tuttavia, ogni volta che un individuo è connesso alla rete (anche tramite sensori) lascia numerose "tracce", che vengono cedute agli operatori online sia in modo informato, sia, più spesso, inconsapevolmente. L'impronta digitale (*online footprint*) di ciascun individuo si compone di numerosissime informazioni, alcune delle quali direttamente associabili allo stesso (nome, cognome, età, ecc.), altre associabili alle attività svolte dagli individui (pagamento di fatture, situazione finanziaria, ecc.), altre, infine, che pur non presentando legami diretti con l'individuo, attraverso il loro processamento, possono facilmente essere associate alle persone; ciò nonostante l'utilizzo di strumenti tecnici mirati all'anonimizzazione dei dati.

Va ricordato, tra l'altro, che le sole informazioni riguardanti il nome e il cognome molto spesso non risultano sufficienti per individuare uno specifico individuo: ad esempio, il nome Mario Rossi non consente di identificare uno specifico individuo, dal momento che si indentificano tutti i Mario Rossi esistenti. Molto spesso, al contrario, l'identificazione di un soggetto specifico avviene tramite l'utilizzo di informazioni alternative che consentono di individuare in maniera univoca una persona.

Diventa, quindi, sempre più rilevante spostare l'attenzione sugli operatori che raccolgono online i dati, sulle modalità di raccolta, processamento e conservazione dei dati, nonché sullo scopo di utilizzo di tali informazioni, anche se in questo caso bisogna tener presente, come descritto in precedenza, che in molti frangenti l'utilità di un dato non è nota al momento della raccolta, ovvero emergono in un secondo momento modalità di utilizzo sconosciute nella fase di acquisizione.

È utile ricordare, inoltre, che lo stesso dato potrebbe essere considerato come "dato personale" nelle mani di un determinato soggetto, mentre potrebbe perdere tale caratteristica nel caso in cui sia in possesso di un altro soggetto. L'esempio che può essere fatto è quello di una foto in un luogo pubblico (ad esempio una manifestazione pubblica): un giornalista potrebbe avere interesse solo a mostrare la folla presente, mentre un pubblico ufficiale potrebbe utilizzare la stessa informazione (la foto) per identificare le persone presenti. Questo fenomeno, tra l'altro, accade sempre più spesso data la sempre più pervasiva presenza

[51] Cfr. Agcom, "Il consumo di servizi di comunicazione: esperienze e prospettive" (pubblicato il 20 ottobre 2016), in particolare, v. Capitolo 3, laddove si evidenzia come "*La rilevanza di internet nella vita quotidiana trova conferma tra i consumatori; l'accesso alla rete, infatti, è ritenuto un servizio indispensabile per oltre il 90% degli individui*" (pag. 22).

di fotocamere che i privati e le amministrazioni pubbliche posizionano sul territorio ai fini della video sorveglianza. Molto, quindi, dipende anche dallo scopo sotteso al processamento dei dati, nonché, come descritto in precedenza, dalla eventuale ricerca di correlazioni spurie e dall'utilizzo sempre più massiccio di algoritmi e tecniche di *machine learning*. In altre parole, **le caratteristiche del dato dipendono anche "dagli occhi di chi guarda".**

Aiutano a formare il complesso di dati digitali **tutte le attività che gli individui svolgono navigando su internet o comunque quando è connesso alla rete** (e in taluni casi anche in modalità temporaneamente *off-line*): l'utilizzo dei motori di ricerca, la lettura delle *news*, la visualizzazione di video, il fare acquisti, il giocare online, ecc., sono tutte attività che lasciano tracce digitali, cioè dati, rilasciati più o meno consapevolmente dagli individui. Il solo fatto di connettere il proprio *device* alla rete di per sé produce un ammontare considerevole di dati digitali, visto che è richiesto un indirizzo IP il quale contiene dati geografici; tutte le pagine che un utente visita formano una scia di dati che viene registrata e dalla quale possono essere desunte numerose informazioni sui gusti (o preferenze) che guidano le scelte di consumo, sugli usi e sulle abitudini dell'utente, spesso senza che l'utente percepisca il fatto di stare rilasciando dati, a fronte di sistemi di *data analytics* che consentono la raccolta e la elaborazione di tali informazioni.

Come evidenziato in precedenza (v. paragrafo 1.1.2), un'altra attività che crea, in maniera costante nel tempo, una grande quantità e varietà di dati digitali proviene dall'**utilizzo della posta elettronica**, uno dei principali strumenti utilizzati, in ambito soprattutto lavorativo, per comunicare; dati strutturati e non, come testo, immagini e la rete di conoscenti, possono essere facilmente raccolti, immagazzinati e utilizzati. Ciò è vero anche per quei software che consentono di comunicare in tempo reale che oltre il testo utilizzano immagini, audio, video, file, ecc.

Chiaramente, una mole rilevante di dati proviene dall'utilizzo di ***social network*** (cfr. Capitolo 3), così ampiamente diffusisi negli ultimi anni tra la popolazione. Anche in questo caso, la consapevolezza degli utenti circa l'utilizzo dei propri dati appare piuttosto ridotta: *posts*, commenti, immagini, contatti di un individuo sono dati che vengono raccolti ed utilizzati, senza che l'utente spesso si renda conto di cederli. Tra l'altro, l'utente ha poca contezza circa i fini di utilizzo dei propri dati.

Altro elemento fondamentale di questo ecosistema riguarda la **mobilità**. Oggi, infatti, gli utenti non sembrano in grado di fare a meno dell'uso dei servizi in mobilità, che consentono in ogni momento la localizzazione dell'apparecchio (*Located-Based Services – LBS*). Come conseguenza, il percorso giornaliero (*routing*) del possessore del *device* viene tracciato con precisione.[52] Inoltre, sempre più spesso gli apparecchi mobili sostituiscono i PC da postazione fissa, cosiddetto *mobilePC* (*mPC*), con le relative attività (posta elettronica, messaggistica, visione di video, *social network*, ecc.) e ciò, come descritto in precedenza (cfr. 1), ha immediatamente generato lo sviluppo di tecniche che consentono il tracciamento "*cross-device*" delle preferenze e delle abitudini degli individui (si pensi, ad esempio, alla visione di un video che inizia da postazione fissa, poi viene interrotta, per poi essere ripresa da un apparecchio mobile).

Infine, i dati digitali relativi alle singole persone sono raccolti sempre più spesso tramite **sensori e sistemi di sensori** che oramai già pervadono la vita quotidiana dei cittadini: ad esempio, i sistemi di video sorveglianza e i pannelli pubblicitari con sensori ottici hanno concorso allo sviluppo di tecniche che consentono il riconoscimento facciale, contribuendo in modo significativo alla crescita della varietà, velocità e volume dei dati digitali. Con il proliferare dei sensori e dell'internet delle cose (IoT) anche l'utilizzo dei beni da parte degli utenti viene costantemente monitorato e registrato. A tal proposito, è utile ricordare che anche i moderni *device* mobili contengono una serie di sensori da cui originano una

[52] Ad esempio, ciascun possessore di uno smartphone può, iscrivendosi ai servizi di *Google*, verificare la precisione con cui vengono raccolte informazioni sulla cronologia degli spostamenti: https://maps.google.com/locationhistory.

molteplicità di rilevanti informazioni (accelerometro, giroscopio, magnetometro, rilevatore di prossimità, lettore delle impronte digitali e facciali, riconoscimento, sensore di luminosità, termometro, GPS, ecc.). Tale evoluzione non potrà che subire un'ulteriore accelerazione con la futura realizzazione delle reti mobili di quinta generazione (cd. 5G).[53]

Le sorgenti di dati digitali strettamente legate agli individui (connessione in rete, utilizzo della posta elettronica, uso dei servizi di telecomunicazioni mobili, sensori e sistemi di sensori) producono un flusso di dati continuo (velocità), molto variegato (varietà) e molto denso (volume). Questo flusso, è bene ricordarlo, solo in parte si compone di dati ceduti consapevolmente dall'utente; una parte sempre più rilevante, infatti, viene raccolta senza il consenso esplicito degli utenti, i quali in maniera passiva (*passive data*) diventano una fonte primaria di informazioni.

La raccolta sempre più pervasiva di dati relativi ad abitudini e preferenze degli individui, unitamente alla capacità, attraverso la loro analisi, di scovare modelli ad oggi sconosciuti, hanno fatto sì che si concretizzassero gigantesche opportunità di innovazione, ma anche considerevoli rischi. Inoltre, la stessa nozione di *big data*, ovvero *big data analytics,* suggerisce che il valore delle informazioni non è legato esclusivamente alla crescente capacità di raccolta dei dati o alla qualità degli stessi, ma soprattutto alla successiva possibilità di compiere, ovvero supportare, attraverso quelle stesse informazioni, processi decisionali (spesso in tempo reale): all'aumento delle informazioni corrisponde la conversione delle stesse in conoscenza (cfr. **Figura 1.7**).

A fronte della rivelazione di proprie informazioni, l'individuo trae **benefici** tangibili ed immediati in termini di compensazioni monetarie sotto forma di sconti o premi, accesso a trattamenti preferenziali ovvero a servizi gratuiti e personalizzati. Di converso, alla data *disclosure* nei confronti di terzi si associano **costi** come quelli di "invasione" che l'individuo deve sostenere per far fronte a spam, telemarketing, e pubblicità (anche via email), fino a veri e propri furti di identità, ma anche costi legati al fatto che, per mezzo della "profilazione" dell'utente, è sempre più probabile ricevere offerte selettive (disegnate cioè sulle proprie intenzioni di spesa) e, pertanto, essere destinatari di politiche di discriminazioni di prezzo.

In un ambiente popolato da pochi *gatekeeper* (v. paragrafo 1.5), accade che poche imprese private, tramite la rilevazione delle preferenze, dei desideri, degli interessi e delle abitudini di milioni di persone, hanno fatto della raccolta e sfruttamento dei dati il proprio *core-business*.

I dati sono il fattore propulsivo di un ecosistema al centro del quale alcune piattaforme a più versanti offrono, da un lato, servizi gratuiti ad individui che, in cambio degli stessi, cedono, più o meno consapevolmente, i propri dati. Sull'altro versante della piattaforma, e grazie ai dati acquisiti presso gli individui, vengono offerti, a tutti gli agenti che vogliano farne uso, servizi di *data analytics*, profilazione di potenziali clienti e spazi pubblicitari mirati.

In tale contesto, quindi, i dati non sono che una "merce di scambio". Nell'era dell'economia digitale, in effetti, è ampiamente diffusa la pratica per cui il vero valore della transazione tra consumatori e imprese non riguarda il consumo di beni e servizi, che, come ricordato, spesso si presentano come gratuiti, ma dallo scambio (per lo più implicito) delle informazioni sottostanti (v. paragrafo 2.7).

Nell'era dei *big data*, i singoli individui contribuiscono, il più delle volte in modo inconsapevole, in maniera determinante alla creazione del flusso di dati digitali; come due facce di una stessa medaglia, da un lato, gli utenti beneficiano di migliori servizi; dall'altro, tuttavia sopportano costi derivanti dalla cessione di informazioni, con un processo redistributivo tipico delle economie digitali.

[53] Cfr. Agcom, Indagine conoscitiva concernente le prospettive di sviluppo dei sistemi wireless e mobili verso la quinta generazione (5G) e l'utilizzo di nuove porzioni di spettro al di sopra dei 6 GHz, pubblicata il 5 marzo 2018.

2.3.Le caratteristiche economiche dei dati

L'interesse degli economisti per il mercato dei dati si è focalizzato principalmente sulla dimensione informativa emersa all'interno dello stesso. L'analisi si è cioè concentrata sui *trade-off* derivanti dalla scelta indicativamente razionale, basata cioè sulla conoscenza dei relativi rischi e benefici, dell'individuo in ordine alla cessione di informazioni che lo riguardano. Se la cessione dei dati da parte di un individuo avviene nell'ottica dell'analisi costi-benefici, è allora possibile parlare dello scambio di dati alla stregua della cessione di una qualsiasi merce, vale a dire di *"**data as a commodity**"*.

In tal senso, diventa innanzitutto rilevante analizzare quali siano le **caratteristiche del bene economico rappresentato dai dati**. Come già sottolineato precedentemente (cfr. paragrafo 1.4), la proliferazione e l'utilizzo di dati a fini commerciali a cui abbiamo assistito negli ultimi anni sono stati paragonati alla scoperta degli idrocarburi e alla loro applicazione all'industria moderna.[54] In effetti, il paragone appare abbastanza condivisibile se si pensa ai dati come *input* produttivo in grado di innescare una rivoluzione commerciale tale da generare nuovi prodotti e servizi che aumentano il benessere dei consumatori e che, allo stesso tempo, presenta taluni rischi (alla stregua dell'inquinamento) in grado di ridurre la portata degli stessi benefici sociali.

I caratteri distintivi dei *big data*, in qualità di beni economici, possono essere sintetizzati facendo ricorso ad alcuni concetti classici utilizzati in ambito economico. In primo luogo, il dato si presenta come un bene **non scarso**. La scarsità rappresenta una caratteristica molto rilevante dei beni economici che si trasferisce in maniera diretta sul livello del prezzo; più un bene è scarso, infatti, maggiore è il prezzo di equilibrio che si formerà sul mercato. Come descritto in precedenza (v. paragrafo 1.1), l'ammontare dei dati non solo presenta una crescita esponenziale nel volume, ma anche nella varietà e soprattutto si sta diffondendo ad una velocità senza precedenti; questa abbondanza di informazioni rende di fatto molto problematica l'individuazione di un valore economico attribuibile al singolo dato. Mentre un barile di petrolio ha un valore intrinseco, al punto da rappresentare l'unità di misura per individuare il prezzo di riferimento del bene sul mercato, un singolo dato di per sé presenta difficilmente un valore economico; è, infatti, attraverso l'aggregazione di più dati e la loro successiva analisi che si è in grado di estrarre valore dai dati stessi.

In secondo luogo, il dato si presenta come un bene caratterizzato da **non rivalità** nel consumo, caratteristica che in parte permette di accostare i dati ad un bene pubblico;[55] in effetti, l'uso di dati da parte di un agente non incide sulla facoltà di goderne completamente da parte di terzi, ovvero in altri termini, lo stesso dato può essere riutilizzato senza che il suo riutilizzo ne determini una riduzione di valore. Inoltre, i dati presentano in parte anche la caratteristica della **non escludibilità**, vale a dire l'impossibilità di estromettere terzi dal loro consumo; nonostante l'interesse e la presenza di norme a tutela della riservatezza e dell'integrità dei dati relativi agli individui, la condizione di non escludibilità non sempre si realizza a pieno (v. in tal senso anche il paragrafo 2.3). Da un lato, infatti, lo sviluppo tecnologico consente in maniera estremamente facile e a costi marginali prossimi a zero di riprodurre lo stesso dato più volte; ciò riduce sensibilmente la possibilità di esclusione di scambi di dati. Dall'altro, più complessi sono i dati raccolti, più è possibile rendere i dati esclusivi, creando di fatto barriere commerciali al loro scambio e, quindi, ad un loro uso diffuso.[56]

---

[54] *The world's most valuable resource is no longer oil, but data*, Economist, 2017; *Why data is the new oil*, Fortune, 2016; *From fintech to techfin: data is the new oil,* The Asian Banker, 2016.

[55] Un bene si definisce "bene pubblico puro" quando tutti possono consumare *simultaneamente* (non rivalità nel consumo) lo stesso bene e *nessuno può essere escluso* (non escludibilità nello scambio) dal consumo di quel bene. È bene ricordare che nella realtà le due caratteristiche si combinano in maniera tale da fornire uno scenario molto variegato di beni pubblici, a seconda della maggiore o minore intensità con cui si presentano le due caratteristiche.

[56] VAN TIL H., VAN GROP N., PRICE, K., (2017), *Big data and Competition Policy*, Ecoryse per conto del Ministero dell'economia Olandese.

L'analisi economica utilizza i concetti di **sostituibilità** e **complementarità** dei beni (o dei fattori produttivi) rispetto alle preferenze nei gusti (o nella combinazione dei fattori produttivi); in tal senso, e in maniera intuitiva, i dati possono presentare entrambe le caratteristiche, anche se il mondo dei *big data* si basa in modo preponderante sul concetto di complementarità d'uso. Infatti, come più volte sottolineato, la complementarità tra dati, di diverso formato e da diversa fonte, risulta preponderante: la capacità di aggregare fonti eterogenee di informazioni risulta fondamentale per estrarre valore dai dati. Tuttavia, tra i dati è presente anche un certo grado di sostituibilità; ad esempio, ai fini della realizzazione di una campagna di marketing, i dati sui comportamenti dei consumatori relativi alle scelte di consumo possono essere in parte sostituiti dai dati sul consumo effettivo di beni e servizi; ma anche in questo caso è dalla combinazione di entrambi i tipi di dati che si possono ottenere informazioni ancora più puntuali in grado di rendere la campagna pubblicitaria più efficace.[57]

In linea generale, i beni economici tipicamente considerati nella letteratura possono essere classificati anche secondo la velocità con la quale perdono di valore, vale a dire la loro **deperibilità**. Anche per tale caratteristica la complessità dei *big data* rende di fatto difficile una classificazione dei dati in base alla loro deperibilità. Alcuni dati, infatti, perdono di valore quasi immediatamente dopo il loro utilizzo; si pensi ai dati sulle condizioni relative alla viabilità che possono avere un valore enorme in un momento specifico e non valere quasi più nulla negli attimi successivi. Tuttavia, in virtù della possibilità di riutilizzo dei dati, nonché del loro valore opzionale, la perdita di valore di un *asset* così strategico, alla stregua degli altri *asset* patrimoniali, avviene molto lentamente. Infatti, la caratteristica dei *big data* è quella di fornire un valore informatico dinamico di tipo collettivo.

Ciò vuol dire che il singolo dato non solo e non tanto fornisce informazioni puntuali sull'individuo, ma permette anche di individuare pattern comportamentali sociali, che poi vengono sfruttati, anche successivamente e nelle maniere più svariate per estrarre valore. In tal senso, **la somma dei valori dei singoli dati presi in un dato istante è assai diversa (e decisamente inferiore) al valore dei dati presi nel loro complesso e in un più ampio intervallo temporale**. Tale caratteristica, che ha implicazioni anche in termini di struttura dei costi delle imprese (elevati costi fissi e affondati e costi marginali prossimi a zero), non può che condurre a fenomeni di concentrazione dei mercati.

La sintetica analisi delle caratteristiche economiche dei dati è particolarmente interessante in quanto evidenzia come non sempre l'applicazione di concetti appartenenti al paradigma economico classico siano utili a spiegare un fenomeno così complesso come i *big data*. In particolare, l'idea di considerare i dati come un bene tradizionale non consente, ad esempio, l'uso dei tradizionali strumenti di contabilità nazionale utilizzati per misurare i flussi di scambio di dati tra i vari paesi, così come avviene per qualsiasi altro bene.[58] La fondamentale identità del commercio internazionale, in base alla quale i consumi domestici sono uguali alla produzione domestica più le importazioni a cui bisogna sottrarre le esportazioni, ben si adatta a beni come gli idrocarburi o alle classiche materie prime, ma presenta enormi difficoltà applicative in riferimento ai dati. Basti pensare al mercato delle applicazioni, le cosiddette APP, per i *device* mobili (v. paragrafo 2.5); la maggior parte degli applicativi, infatti, è presente contemporaneamente su più mercati geografici senza che ciò implichi un processo di esportazione del servizio stesso. In effetti, se un paese vuole aumentare le esportazioni di un bene, a parità di produzione, dovrà ridurne il consumo interno; quando si tratta di dati, invece, le esportazioni non riducono l'ammontare di dati disponibili internamente, in virtù anche della non rivalità nel consumo.

[57] VAN TIL H., VAN GROP N., PRICE, K., (2017), *Big data and Competition Policy*, Ecoryse per conto del Ministero dell'economia Olandese.

[58] "La Contabilità Nazionale (CN) è la descrizione quantitativa dell'attività economica di un Paese, sotto forma di una completa e sistematica presentazione dei flussi economici e finanziari che si verificano tra gruppi significativi di operatori e delle consistenze finali dei beni reali e finanziari", SIESTO V., (2003), *La contabilità nazionale*, Il Mulino.

Un passo successivo nell'analisi economica dei *big data* concerne l'individuazione di possibili ***trade off*** **che avvengono nel momento in cui i singoli individui sono chiamati a prendere decisioni in ordine alla cessione o meno di informazioni**. L'individuo che decide di prestare il consenso alla cessione dei propri dati compie, a fronte di una conoscenza parziale dell'ambiente in cui la valutazione è compiuta e, quindi, dei rischi e dei benefici che ne derivano, una valutazione in termini di costi-benefici analoga a quella che fronteggia ogniqualvolta deve decidere se effettuare o meno un acquisto. Più nello specifico, l'individuo compie una valutazione orientata a comprendere se sia conveniente cedere le proprie informazioni in cambio di determinati benefici, anche se non sempre sono benefici suscettibili di valutazione economica.

Si tratta di uno scambio che avviene sulla base della presenza di significative e strutturali **asimmetrie informative** tra gli agenti in causa (in questo caso, tra gli utenti che cedono dati e gli operatori che li acquisiscono e li utilizzano); ciò si tramuta per gli individui in un contesto nel quale, non avendo accesso a tutte le informazioni, è di fatto impossibile procedere a una corretta misurazione dei costi (incerti e potenziali).[59] Gli individui non dispongono della stessa quantità e qualità di informazioni in possesso di chi le raccoglie e mette insieme e che, quindi, fa gravare sugli individui stessi un costo rappresentato dalla *disclousure* delle informazioni individuali. Dell'utilizzo che di questi dati verrà fatto, inoltre, gli individui (fonte originaria del dato) non sono a conoscenza, e anzi, i benefici derivanti da tale tipologia di transazione non li vede per nulla coinvolti.

Le scelte degli individui, quindi, sono compiute in un ambiente caratterizzato da componenti esogene ed endogene di difficile valutazione, tra le quali emergono, con forza, la presenza di incertezza e di dipendenza dal contesto (*context dependence*).[60]

L'**incertezza** è in questo caso funzione del progresso tecnologico e dell'attività di data *collection*; l'individuo, infatti, dovrà valutare il fatto che la tecnologia celi, da un lato, la tipologia e la modalità di raccolta dei dati e, dall'altro, il loro uso successivo. È raro che gli individui abbiano realmente contezza di quali e quante informazioni vengano acquisite dagli operatori online con cui entrano, sempre più di frequente, in contatto, ma soprattutto è molto complicato risalire alle tipologie di impiego da parte di chi raccoglie i dati. La difficoltà di comprendere quali siano le conseguenze dell'eventuale cessione di dati fa sì che l'analisi dei costi e dei benefici si connoti di un significativo grado di incertezza.

Un altro fattore che determina l'intensità delle asimmetrie informative e aumenta l'incertezza riguarda la dipendenza della scelta dallo specifico contesto (***context dependence***); tale componente fa sì che lo stesso soggetto possa esplicitare preferenze diametralmente opposte rispetto alla scelta di cedere i propri dati in funzione, per l'appunto, del contesto: al variare delle "condizioni ambientali", gli individui passano cioè da un atteggiamento di profonda diffidenza rispetto all'eventualità di cedere i propri dati, ad una totale indifferenza rispetto alle conseguenze di una simile scelta. Data la presenza di asimmetrie informative, inoltre, non di rado chi è interessato alla raccolta di dati utilizza ogni possibile strategia per favorire atteggiamenti che prevedono la cessione di dati. In pratica, la dipendenza dal contesto non fa che aumentare ulteriormente l'incertezza sistemica.

In conclusione, le scelte di un individuo in ordine alla cessione di propri dati al fine di ottenere un servizio si indirizzano a seconda del bilanciamento operato tra benefici, spesso immediati (es. l'accesso ad un servizio), e costi (spesso incerti e non conosciuti). In questo contesto, l'asimmetria informativa tra utenti e operatori è pervasiva e strutturale: non solo il consumatore non ha a disposizione tutte le

[59] VARIAN H.R., (1999), *Economic aspects of Personal Privacy*, in Privacy and self-regulation in the information age, National Telecommunications and Information Administration.

[60] ACQUISTI A., BRANDIMARTE L., LOEWENSTEIN G., (2015), *Privacy and Human behavior in the age of information*, Science 347 n. 6221.

informazioni di cui avrebbe bisogno per prendere una scelta informata e razionale, ma molti dei comportamenti, per essere efficienti, presupporrebbero un grado di conoscenza tecnica che va molto al di là delle competenze diffuse tra la popolazione. In altre parole, **un maggior grado di trasparenza risulta in molti casi inutile laddove i consumatori non riescano, a causa di uno strutturale gap di conoscenze tecnologiche, a elaborare correttamente tali informazioni.**

La letteratura ha inoltre dimostrato come il comportamento umano, specie in condizioni di incertezza, non sia affatto razionale. **Scelte, come quelle relative alla cessione dei propri dati, vengono effettuate assai frequentemente di impulso e senza una valutazione delle reali conseguenze dello scambio implicito**[61].

Questi elementi hanno poi effetti sull'**efficienza complessiva circa il funzionamento dei mercati**. La letteratura economica in materia non fornisce risposte univoche alla domanda su quale sia il livello ottimale di *disclousure* dei dati legati ai singoli individui. Secondo l'approccio legato alla cosiddetta *Scuola di Chicago*, un'eccessiva protezione dei dati si tramuta in una riduzione dell'efficienza dei mercati, dal momento che le imprese non ricevono segnali a sufficienza su come allocare in maniera efficiente i propri input produttivi. In altri termini, un eccesso di restrizioni determinerebbe un trasferimento dei costi dagli individui alle imprese che quindi si troverebbero ad operare in maniera meno efficiente, sia da un punto di vista statico (efficienza legata all'allocazione delle risorse), che da quello dinamico (efficienza legata alle innovazioni).[62] D'altra parte, tale tipologia di scambio esula dalle logiche di mercato, trattandosi non di semplici merci ma di aspetti riguardanti diritti fondamentali dell'uomo, sia individuali che collettivi. Se molta dell'analisi che viene effettuata in questo Rapporto si riferisce proprio alla tutela di questi diritti, e in particolare a quello costituzionale all'informazione (v. in particolare Capitolo 3), in questa sede si analizzano in maggior dettaglio gli aspetti economici.

Dalle argomentazioni che precedono, **risulta evidente come lo scambio di dati dia spesso luogo a strutturali fallimenti di mercato**. In primo luogo, come nel caso delle citate emissioni di idrocarburi, gli investimenti posti in essere dalle imprese per la raccolta di dati sugli individui, non internalizzando i costi sociali, rischiano di condurre ad una situazione di sovrainvestimento nella raccolta delle informazioni (nel caso degli idrocarburi, una sovrapproduzione di inquinamento)[63].

Come descritto poc'anzi, molto rilevante in questo dibattito è l'effetto prodotto dal contesto; ad esempio, è chiaro che di fronte alla possibilità di ottenere nell'immediato sconti o servizi gratuiti e/o personalizzati, l'individuo sarà portato a rilasciare dati individuali come quelli relativi ai propri gusti e preferenze senza considerare i costi derivanti dalla loro divulgazione.[64] In un contesto in cui sono presenti costi di transazione e incertezza riguardo la corretta assegnazione dei diritti di proprietà sui dati (ad esempio a chi spetta il diritto nel caso di rivendita del dato a terze parti, oppure come sempre più spesso accade, a chi spetta il diritto di proprietà nel momento in cui i dati legati ad un individuo vengono aggregati con altri dati), è assai probabile che le forze di mercato non siano in grado di garantire il raggiungimento di una situazione di efficienza dell'equilibrio economico. Si concretizza, in particolare, la possibilità che a prevalere siano gli interessi di coloro che detengono maggiori conoscenze tecniche e informazioni riguardo ai dati stessi.

[61] KAHNEMAN D., TVERSKY A. (1979), *Prospect Theory: An Analysis of Decision under Risk*, Econometrica 47, n. 2, pp. 263-292.

[62] STIGLER G.J., (1980), *An introduction to privacy in economics and politics*, Journal of Legal Studies 9, n.4; POSNER R.A., (1981), *The economics of privacy*, American Economic Review 71, n. 2.

[63] HIRSHLEIFER J., (1980), *Privacy: its origin, function and future*, The Journal of Legal Studies 9, n. 4.

[64] VARIAN H.R., (1996), *Economic aspects of Personal Privacy*, in Privacy and self-regulation in the information age, National Telecommunications and Information Administration.

2.4. Le strategie di discriminazione

La crescita della *datasphere* (v. paragrafo 1.1.1) porta con sé grandi opportunità di crescita e sviluppo, non solo di natura economica, strettamente collegata alla nascita di nuovi prodotti e servizi o al miglioramento di quelli già esistenti, ma anche di natura più prettamente sociale, laddove, ad esempio, si riscontrano miglioramenti nel campo della medicina e nella gestione della cosa pubblica. Tuttavia, un numero crescente di analisi, evidenzia il sorgere di problemi connessi a **forme di discriminazione**, tra cui anche la disparità nell'accesso ai dati e agli strumenti utilizzati per analizzarli.

Oltre alla ricerca di nuove opportunità di *business*, le imprese utilizzano i *big data* allo scopo non solo di predire i consumi futuri, ma anche di orientarli verso un prodotto piuttosto che un altro. Tali processi sono oggi resi possibili dall'utilizzo sempre più massiccio di **algoritmi attraverso cui indirizzare le scelte degli individui**[65]. Ciò accade, ad esempio, quando comunichiamo con amici e familiari, quando scegliamo un percorso stradale, quando effettuiamo ricerche sui motori di ricerca, quando ci informiamo sui *social network*. A questo ultimo riguardo, secondo un recente studio, la maggioranza degli utenti dei *social*, in particolare tra coloro non in possesso di un background informatico, non è a conoscenza del fatto che le notizie che appaiono al consumatore (cosiddette *news feed*) siano filtrate da un algoritmo.[66]

Tra le possibili **pratiche discriminatorie**, quelle **legate al prezzo** sono tra le più diffuse; il prezzo, infatti, è una delle variabili più importanti attraverso cui le imprese cercano di realizzare la massimizzazione dei loro profitti quando la configurazione di mercato differisce da quella concorrenziale, ovvero quando le imprese possono esercitare il loro potere di mercato. La discriminazione di prezzo è quella pratica che consente all'impresa di offrire lo stesso bene o servizio a prezzi differenti (discriminazione nel prezzo) a seconda della disponibilità a pagare dei singoli consumatori (o prezzo di riserva), così come individuata dagli operatori tramite tecniche di *big data analytics*.[67]

La discriminazione di prezzo rappresenta un fenomeno molto rilevante soprattutto in seguito alla diffusione degli acquisti online effettuati dai consumatori. Esistono diversi modi attraverso cui le imprese riescono a vendere uno stesso bene o servizio a prezzi differenti; una delle classificazioni che più comunemente viene utilizzata in ambito economico, e che ben si attaglia alle problematiche dei *big data*, si basa sul livello di informazioni in possesso delle imprese riguardo i gusti e le preferenze dei consumatori, dal momento che queste informazioni risultano fortemente correlate con la disponibilità a pagare dei medesimi.

A seconda della tipologia di informazione a disposizione dell'impresa è possibile classificare la discriminazione di prezzo in tre tipologie: 1°, 2° e 3° grado. Quella di 3° grado presuppone che solo alcune caratteristiche degli acquirenti siano ben osservabili e, quindi, queste informazioni possano essere utilizzate per predisporre tariffe diverse in funzione delle stesse; si tratta di strategie spesso utilizzate quando è possibile fissare prezzi differenti in base a caratteristiche facilmente osservabili quali età, genere e talvolta occupazione (si pensi a prezzi scontati per fasce di consumatori più giovani o mature, a ingressi gratuiti a seconda del genere, e a biglietti scontati per categorie come gli studenti). Per attuare strategie di

[65] Per una illustrazione del ruolo degli algoritmi nel mondo dell'informazione, cfr. Box 5.1.
[66] HAMILTON K., SANDVIG C., KARAHALIOS K., ESLAMI M., (2014), *A Path to Understanding the Effects of Algorithm Awareness*, ACM, Toronto.
[67] È utile ricordare che in un mercato perfettamente competitivo vale la legge del prezzo unico; ciò è conseguenza del fatto che se fosse possibile discriminare si metterebbero in moto dei meccanismi di arbitraggio tali per cui si verrebbero a generare mercati secondari. Affinché sia possibile attuare, da parte delle imprese, strategie di discriminazione di prezzo, quindi, è necessaria la presenza di potere di mercato, cioè forme di mercato non concorrenziali, e l'assenza di mercati secondari perché i consumatori non sono a conoscenza del fatto che il bene è venduto a prezzi differenti, o perché la rivendita è vietata per legge, come accade per i servizi di energia elettrica. Inoltre, si potrà parlare di discriminazione dei prezzi solo se la differenza nel prezzo per lo stesso bene non possa essere attribuita ad una differenza nei costi di produzione.

discriminazione di prezzo di 2° grado, invece, le imprese utilizzano informazioni relative alla quantità di bene o servizio che i consumatori utilizzano; rilevando l'eterogeneità nei modelli di consumo, infatti, le imprese sono in grado di offrire un ventaglio di offerte rispetto alle quali i consumatori si auto-selezionano, come tipicamente avviene nei contratti di fornitura dei servizi di telecomunicazione (tariffe a due stadi), nelle offerte "paghi due e prendi tre", ovvero in tutte quelle strategie di marketing volte alla fidelizzazione (sconti quantità, tessere fedeltà, ecc.). Infine, la discriminazione di 1° grado, anche detta "perfetta discriminazione", si ha quando il venditore è in grado di applicare, a ciascun cliente, un prezzo corrispondente alla stima del prezzo massimo che lo stesso è disposto a pagare; si tratta di un caso che in economia è stato considerato come meramente teorico, in considerazione dell'elevato set informativo di cui le imprese dovrebbero disporre per attuare una simile strategia. **L'enorme disponibilità di dati individuali connessa all'avvento dei *big data* sta rendendo sempre più concreta la possibilità per gli operatori online di attuare strategie di perfetta discriminazione di prezzo.**

In linea generale, le strategie di discriminazione del prezzo sono considerate come un fenomeno in grado di accrescere, o comunque non diminuire, il benessere sociale e, quindi, per certi versi sono state in passato giudicate auspicabili. Attraverso tali pratiche, infatti, da un lato si può aumentare il benessere sociale perché è possibile realizzare la vendita di beni che altrimenti non sarebbero mai stati venduti in assenza dell'individuazione, da parte del venditore, di consumatori disposti a pagare un prezzo più alto rispetto ad altri; dall'altro lato, si riesce a garantire una maggiore partecipazione allo scambio tramite il coinvolgimento anche di quei consumatori che, presentando una bassa disponibilità a pagare, senza discriminazione non avrebbero partecipato allo scambio dal momento che avrebbero dovuto sostenere un prezzo effettivo superiore rispetto a quello massimo che sono disposti a pagare.

La pratica della discriminazione di prezzo, tuttavia, è stata, in alcuni casi, contestata facendo leva su una questione di equità piuttosto che su basi puramente economiche; in maniera intuitiva, infatti, offrire uno stesso bene a prezzi differenti provoca un danno a coloro che mostrano una disponibilità a pagare superiore. Se la discriminazione è attuata in modo efficace, inoltre, può permettere all'impresa di raggiungere il massimo profitto disponibile nel mercato, lasciando i consumatori senza nessun guadagno dalle transazioni.[68] Infine, è utile evidenziare anche l'esistenza di alcune criticità legate ai costi sostenuti dalle imprese per porre in essere la discriminazione di prezzo, dal momento che simili strategie necessitano di allocare risorse che altrimenti sarebbero destinate ad altre attività.[69]

L'avvento dei *big data* ha reso sempre più frequente l'applicazione di strategie da parte delle imprese che prevedono l'offerta di uno stesso bene o servizio a prezzi differenti; la possibilità di realizzare forme di discriminazione di 1° grado, come ricordato poc'anzi, deriva principalmente dalla possibilità di utilizzare le informazioni digitali riguardanti un singolo individuo, che rappresentano una buona base informativa, per inferire riguardo la sua disponibilità a pagare. Peraltro, **grazie all'individuazione di pattern generali attraverso tecniche di *big data analytics*, occorrono poche informazione individuali per prevedere il comportamento economico e sociale di ciascun individuo** (ad esempio, sono sufficienti pochi *like* su un *social network* per prevedere, con una elevatissima probabilità di successo, informazioni sensibili quali quelle relative a orientamento politico, credo religioso, etnia, orientamento sessuale, situazione sentimentale e, addirittura, dipendenza da sostanze stupefacenti; cfr. Box 1). Ciò rende sempre più praticate le strategie di differenziazione dei prezzi dei beni e servizi venduti online a seconda delle caratteristiche degli individui.

[68] In termini economici tale situazione si riferisce ai casi in cui si riduce la perdita secca di benessere, causata dalla presenza di potere di mercato da parte dell'imprese (come ad esempio il monopolio), e quindi si genera ricchezza aggiuntiva che, però, viene redistribuita esclusivamente alle imprese.

[69] LEESON P.T., SOBEL R.S., (2007), *Costly price discrimination*, Economics Letters 99.

Uno dei primi casi di discriminazione di prezzo, realizzato sulla base delle tracce lasciate in rete dagli individui, ha riguardato, nel settembre del 2000, un cliente di *Amazon.com*, che aveva acquistato un DVD a 24,49$; la settimana successiva, egli notò che lo stesso prodotto aveva subito un rialzo del prezzo di 1,75$ arrivando a costare 26,24$. Lo stesso consumatore riscontrò come la semplice eliminazione dei *cookies* e dei *tag* elettronici dal proprio computer, elementi che facilitano il tracciamento delle attività compiute dagli individui sul web, fosse sufficiente a far tornare il prezzo del DVD al di sotto del livello iniziale, ossia a 22,74$.[70] Tale forma di discriminazione del prezzo è stato uno dei primi casi in cui un'azienda ha posto in essere una strategia di "prezzi dinamici" basata sulla misurazione del desiderio dell'acquirente.[71]

Un ulteriore caso ha coinvolto, nel 2009, l'azienda Microsoft e il suo servizio *Bing Cashback*: pur essendo stato ideato come servizio rivolto ai consumatori affinché potessero ottenere un risparmio nelle loro transazioni online, sembra che alcuni siti di venditori terzi, utilizzando l'URL di provenienza dei consumatori, praticassero forme di discriminazione di prezzo a svantaggio dei visitatori reindirizzati da Bing. In maniera analoga, nel 2012 il sito dell'agenzia di viaggio *Orbitz Worldwide Inc.* ha scoperto che gli utilizzatori dei computer *Mac* facevano registrare in media una propensione alla spesa superiore del 30% per una notta di pernottamento in albergo; di conseguenza, l'agenzia online decise di procedere in primo luogo ad una differente visualizzazione delle offerte, per poi fissare prezzi per utenti *Mac* superiori rispetto agli utenti che utilizzavano *device* "*windows-based*".[72]

Più in generale, l'affermazione dell'*e-commerce* (le transazioni in rete per l'acquisizione di beni e servizi), congiuntamente all'aumento della capacità di aggregazione e di analisi di un'ingente mole di dati - spesso destrutturati e relativi ai consumi degli individui, hanno fatto sì che si moltiplicassero i database contenenti informazioni riguardo alle preferenze di consumo degli stessi. Data l'evoluzione tecnologica, le imprese intendono utilizzare tali database oltre che per approssimare al meglio le propensioni di acquisto degli individui e proporre loro offerte personalizzate, anche per orientare/indirizzare i consumi verso un prodotto piuttosto che un altro, alla ricerca sempre del massimo profitto.

Le pratiche discriminatorie possono essere frutto dell'inconsapevolezza, ma anche di una specifica scelta. A tal proposito, un buon esempio di questa ambiguità è fornito dall'applicazione dei *big data* al contesto del mondo del lavoro. Nel mondo del lavoro si è pensato che con l'avvento dei *big data* fosse possibile ridurre al minimo la soggettività nel reclutamento del personale, attività che porta con sé un certo grado di discriminazione; basando le scelte di reclutamento del personale su un algoritmo che utilizza una serie numerosissima di informazioni sugli individui, sarebbe stato possibile abbattere il *bias* legato alla soggettività nella scelta. Inoltre, l'utilizzo di dati sul comportamento di chi cerca lavoro da fonti come i siti internet potrebbe servire a ridurre anche un tipo di pregiudizio nelle pratiche di reclutamento definito "*social network segregation*" derivante dal fatto che una parte rilevante dei posti di lavoro, in particolare da quando si sono diffusi i *social network*, avviene tramite il passaparola all'interno della propria rete di conoscenze. L'effetto a catena, in maniera intuitiva, è che nelle organizzazioni in cui le minoranze sono già poco rappresentate, le minoranze stesse avranno sempre meno possibilità di accesso ai posti di lavoro e, quindi, saranno sempre più discriminate.

Tuttavia, questi fenomeni discriminatori non sembrano essersi allentati neanche con l'utilizzo delle tecniche predittive dei *big data*. Se gli algoritmi che vengono utilizzati per prendere decisioni in merito al

[70] STREITFELD D., (2000), *On the Web, Price Tags* Blur, The Washington Post.
[71] KRUGMAN P., (2000), *Reckonings; what price Fairness?,* New York Times.
[72] MATTIOLI D., (2012), *On Orbitz, Mac Users Steered to Pricier Hotels*, The Wall Street Journal., WHITE M.C., (2012), *Orbitz Shows Higher Prices to Mac Users*, Time.

personale da assumere non vengono disegnati in maniera ottimale, vi è il forte rischio che tali discriminazioni non solo si perpetuino nel tempo, ma si rafforzino di intensità.[73]

In conclusione, l'accesso a set informativi sempre più vasti rende possibile attuare strategie di discriminazione di prezzo sempre più fini, fino a quelle di perfetta discriminazione. È bene evidenziare come con la individuazione, tramite grandi moli di dati, di specifici pattern comportamentali occorrono poche informazioni per prevedere, con un elevato grado di successo, alcuni comportamenti economici e sociali e specifiche caratteristiche individuali, anche partendo da dati anonimizzati. Le strategie di discriminazione di prezzo comportano un sicuro effetto di redistribuzione sociale a favore degli operatori online. Inoltre, tali pratiche, anche quando efficienti, **presentano rischi sociali molto significativi**. È infatti assai facile comprendere come siffatte strategie algoritmiche possano automaticamente estendersi, anche in modo involontario, a differenze nella popolazione in base a etnia, razza, orientamento sessuale, stato di salute, ecc.

BOX 1 – PSICOMETRIA: IL PROCESSO DI PROFILAZIONE

Negli anni sessanta, due psicologi, Ernest C. Tupes e Raymond Christal, teorizzarono come la personalità degli individui potesse essere tratteggiata attraverso l'individuazione di alcuni elementi essenziali della stessa.[1] Il modello inizialmente messo a punto venne poi affinato fino ad arrivare, all'inizio degli anni novanta e grazie agli studi compiuti da Mccrae & Costa[2] e John[3], all'isolamento di cinque variabili qualitative essenziali descrittive della personalità, note come *"big five"*. Ciascuna di esse era associata e si articolava anche in un carattere speculare che le faceva da contraltare: all'estroversione corrispondeva così l'introversione, alla gradevolezza la sgradevolezza, alla coscienziosità la negligenza, alla nevrosi la stabilità emotiva, all'apertura mentale la chiusura mentale. È possibile far riferimento a tali cinque componenti della personalità anche con l'acronimo OCEAN (**Figura 1.1**): *openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism*.

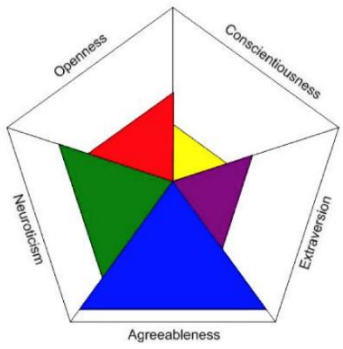


Figura 1.1 – Il modello OCEAN

Fonte: Jennifer Golbeck et al., "Predicting Personality from Twitter," in 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (IEEE, 2011)

[73] MCILVAINE A.R., (2014), *The Power (and Peril) of Predictive Analytics*, Human Resource Executive Online.

Negli anni, si sono susseguite ricerche che hanno provato come la validità del modello fondato sui *"big five"*, ossia basato sulla misurazione psicolinguistica della personalità,[4] sia tuttora adeguato: diversi test, sia in ambito psicolinguistico che psicometrico, non sono infatti stati in grado di metterne in discussione la validità.[5] I *"big five"* sono dunque alla base dell'odierna psicometria: alcune ricerche mostrano come, ad esempio, i tratti della personalità siano deducibili dalle scelte che gli utenti iscritti a *Facebook* compiono nel momento in cui scelgono chi far entrare nella propria rete di contatti (amicizie).[6]

Tramite l'uso di tecniche di *machine learning* e tecniche computazionali appartenenti al mondo dei Big Data, fu Michael Kosinski, creatore dell'app *Mypersonality* – oggi finita al centro dello scandalo che ha coinvolto *Facebook* e *Cambridge Analytica*[7] – a sostenere come "le valutazioni riguardanti la personalità siano più accurate se operate tramite un computer di quanto lo siano se avanzate da un individuo".[8] In particolare, nei suoi studi l'autore ha confrontato l'accuratezza dei giudizi espressi circa la personalità degli individui con i giudizi messi a punto dalle macchine.

Recenti risultati di tale gruppo di studi mostrano che bastano poche decine di *like* per identificare, con una probabilità dell'85%, l'orientamento politico di un soggetto, e il credo religioso con una probabilità dell'82% dei casi (distinguendo tra cristiani e musulmani). Il genere viene invece previsto correttamente nel 93% dei casi (**Figura 1.2**).

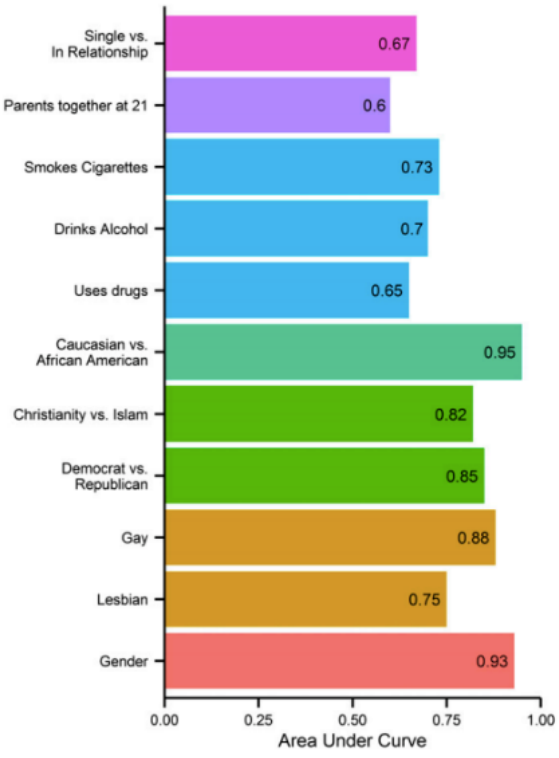


Figura 1.2 – Le predizioni dei modelli
Fonte: Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior"

Bibliografia

[1] Ernest C. Tupes and Raymond E. Christal, "Recurrent Personality Factors Based on Trait Ratings," Journal of Personality 60, no. 2 (June 1, 1992): 225–51, doi:10.1111/j.1467-6494.1992.tb00973.x.

[2] Robert R. McCrae and Oliver P. John, "An Introduction to the Five-Factor Model and Its Applications," Journal of Personality 60, no. 2 (June 1, 1992): 175–215, doi:10.1111/j.1467-6494.1992.tb00970.x. R. R. McCrae and Paul T. Costa, Personality in Adulthood: Emerging Lives, Enduring Dispositions (New York: Guilford, 1990).

[3] OP John, E Donahue, and R Kentle, "Big Five". Factor Taxonomy: Dimensions of Personality in the Natural Language and in Questionnaires," Handbook of Personality: Theory and Research, 1990, 66–100, http://thenetworktufh.org/wp-content/uploads/2015/10/Newsletter2004-02_0.pdf#page=25.

[4] Boele De Raad, "The Big Five Personality Factors: The Psycholexical Approach to Personality," Hogrefe & Huber Publishers, 2000, http://psycnet.apa.org/record/2001-17509-000.

[5] Jennifer Golbeck et al., "Predicting Personality from Twitter," in 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (IEEE, 2011), 149–56, doi:10.1109/PASSAT/SocialCom.2011.33.

[6] Maarten Selfhout et al., "Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach," Journal of Personality 78, no. 2 (April 1, 2010): 509–38, doi:10.1111/j.1467-6494.2010.00625.x.

[7] Carole Cadwalladr and Emma Graham-Harrison, "Revealed: 50 Million *Facebook* Profiles Harvested for Cambridge Analytica in Major Data Breach," The Guardian, 2018, https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html Matthew Rosenberg, Nicholas Confessore, and Carole Cadwalladr, "How Trump Consultants Exploited the *Facebook* Data of Millions," The New York Times, 2018.

[8] Wu Youyou, Michal Kosinski, and David Stillwell, "Computer-Based Personality Judgments Are More Accurate than Those Made by Humans.," Proceedings of the National Academy of Sciences of the United States of America 112, no. 4 (January 27, 2015): 1036–40, doi:10.1073/pnas.1418680112.

2.5. Il mercato delle APP

Uno dei principali trend, che rende i problemi connessi alla gestione dei dati digitali sempre più rilevante, è la rapida crescita della **fruizione dei contenuti su internet tramite *device* mobili**.[74] Tali dinamiche nei consumi sono rese possibili dalla pervasiva diffusione degli apparecchi *handset* che consentono la navigazione sul web, in particolare degli smartphone. A livello globale, **Figura 2.1**, ad ottobre 2016, per la prima volta gli accessi ad internet da postazione mobile (smartphone e tablet su tutti) hanno superato gli accessi da desktop (PC fisso e portatile), stimolati soprattutto dall'uso massiccio di *device* mobili che si registra nei paesi asiatici dove, infatti, gli accessi da postazione mobile hanno superato quelli da postazione fissa già a maggio 2014.

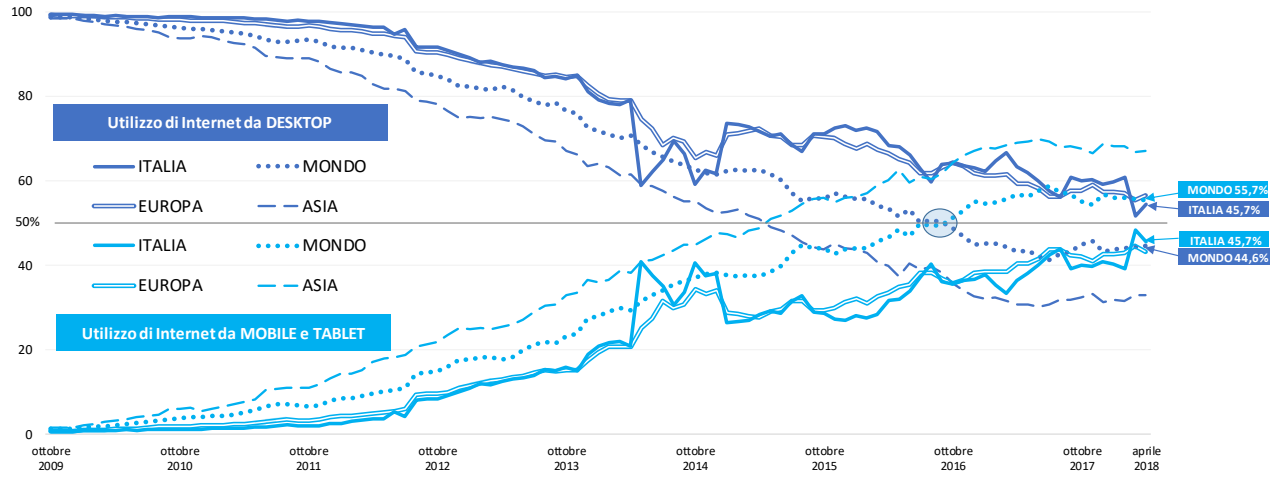


Figura 2.1 – Utilizzo di internet nel mondo per tipologia di dispositivo (ottobre 2009 – aprile 2018)

Fonte: Elaborazioni AGCOM su dati mensili *StatCounter.com*

Le implicazioni di questo *trend* sono molteplici e riguardano vari ambiti; in riferimento a quella che viene definita come "la scia digitale" (*online footprint* v. paragrafo 2.2), di cui ciascun individuo lascia traccia a seguito delle sue attività sul web, senza dubbio l'utilizzo di dispositivi mobili ha dato un forte impulso alla crescita della *datasphere*, dal momento che i *device* mobili, molto più delle postazioni fisse, si caratterizzano per il fatto di essere utilizzati in modo esclusivo dal proprietario/possessore. In linea generale, infatti, i dispositivi mobili appartengono a un singolo individuo (sono pertanto personali), mentre le postazioni da rete fissa si prestano a una maggiore condivisione da parte di una pluralità di individui (es. la famiglia). Prerogativa dei *device* mobili, inoltre, è quella di generare dati associabili alla localizzazione dei singoli utenti; informazione, quest'ultima, sempre più rilevante per la fornitura di servizi *ad hoc* per gli utenti (cfr. paragrafo 1.4).

Tramite i dispostivi mobili, gli individui possono usufruire di servizi di comunicazione in modalità *anywhere* e *anytime*, condizione che ha, in parte, sostituito l'accesso a internet da postazione fissa che appare essere soggetto a vincoli maggiori. Inoltre, la connessione da rete fissa avviene sempre più attraverso *device* mobili che si allacciano alla rete in modalità wireless (spesso wifi) per usufruire di servizi, che, sfruttando

[74] Secondo i dati raccolti dall'Osservatorio sulle comunicazioni, mediante il quale, tra le altre cose, l'Autorità monitora i mercati delle telecomunicazioni, nel corso del 2017 il traffico dati da telefonia mobile è aumentato del 56% e, nello stesso periodo, i consumi unitari sono passati da 1,84 a 2,76 Giga/mese ad utenza, con una crescita del 49,5% (dati a dicembre 2017). Il numero di SIM con accesso ad internet è pari a 52,2 milioni di unità, pari al 63,9% dell'intera customer base, quasi il doppio rispetto al 2012, quando questa tipologia di SIM rappresentava il 27,8% del totale SIM. Cfr. https://www.agcom.it/osservatorio-sulle-comunicazioni. n.1/2018

la connettività *wired*, necessitano di velocità di banda sempre maggiori (streaming audiovideo, videoconferenza,…).

La crescita di connessione a internet tramite *device* mobili rappresenta una condizione strutturale di base per la nascita e la successiva affermazione delle APP,[75] cioè di quei programmi, o pacchetti di programmi - *software applicativi* - che ciascun utente installa, o trova già installati, sul proprio *handset* per lo svolgimento di specifiche attività (es. i programmi di videoscrittura, di *editing* fotografico, di giochi, ecc.). Al riguardo, è stato coniato il termine ***"APP economy"*** per indicare sia l'insieme di attività che comprende l'ideazione, la produzione e la distribuzione di applicazioni mobili, sia i diversi attori protagonisti del mercato.[76] Fondamentalmente, le APP sono create per rendere l'esperienza d'uso dei *device* mobili più agevole e piacevole, permettendo ai singoli utenti di accedere a contenuti e servizi in modo facile, in qualsiasi momento e luogo.

Nel giro di pochi anni, l'ecosistema collegato alle APP ha mostrato una crescita esponenziale; tanto che, rientrando nella categoria di *general purpose technologies*, risulta difficile delimitarne con precisione il perimetro nonché la portata degli effetti sull'intero sistema economico e sociale.[77] Il fenomeno è abbastanza recente, dato che il primo APP *store*, creato dalla società Apple, è stato lanciato nel 2008, un anno dopo l'uscita del primo modello di *iPhone*. Dopo solo tre anni dalla nascita del primo APP *store*, erano già disponibili per gli utenti circa un milione di applicativi distribuiti su quattro negozi virtuali. Il giro d'affari generato varia da ricerca a ricerca, anche a seconda della parte dell'ecosistema che viene maggiormente enfatizzata; secondo i calcoli della società di analisi *App Annie*,[78] l'ecosistema delle APP (comprensivo di applicazioni a pagamento, pubblicità e acquisti mobile) valeva oltre 1.300 miliardi di dollari nel 2016, mentre per il 2021 il giro d'affari previsto arriverà a contare oltre 6.300 miliardi di dollari, con una crescita del 385%. Tale previsione è guidata principalmente da due forze: la diffusione di *device* mobili tra la popolazione ma, soprattutto, la crescita del tempo medio speso nell'uso di APP, in media 3 ore al giorno per utente a fine 2017.[79]

Alcune APP risultano pre-installate (*built-in*) nel *device* mobile, dal momento che con il solo sistema operativo risulta difficile far funzionare il dispositivo; tra queste, è bene ricordare, vi sono anche quelle degli APP *store*, tramite le quali gli utenti hanno accesso al negozio virtuale dove è possibile scaricare le applicazioni. È difficile stabilire quali di queste APP pre-installate siano effettivamente necessarie ai fini del funzionamento del *device* mobile, in quanto molto dipende dalle esigenze dei singoli utenti.[80] In alcuni casi, si tratta di APP indesiderate (*unwanted* APP), ma che risulta molto difficile disattivare dal proprio

[75] Rispetto alle APP disponibili su PC-*Desktop*, per i *device* mobili tali programmi si caratterizzano per una maggiore compattezza e semplicità d'uso che ben si conciliano con le limitate risorse hardware dei dispositivi. Le APP devono pertanto essere in grado di affrontare e superare specifiche problematiche legate alla natura dell'*handset* come ad esempio *i)* la limitatezza delle risorse (memoria, *CPU*), *ii)* l'assenza di alimentazione esterna, *iii)* i differenti protocolli di trasferimento dati per l'accesso a internet (*WiFi*, *GPRS*, ecc.) e *iv)* le ridotte dimensioni del *display*, che rendono molto più critico lo sviluppo dei sistemi operativi.

[76] Nel prospetto informativo finalizzato alla quotazione al NASDAQ nel 2012, la società ZYNGA, specializzata nella produzione di video giochi, tra le opportunità per la crescita del proprio *business* indicava l'emergere della APP economy definendola come segue: "*Emergence of the App Economy. In order to provide users with a wider range of engaging experiences, social networks and mobile operating systems have opened their platforms to developers, transforming the creation, distribution and consumption of digital content. We refer to this as the "App Economy." In the App Economy, developers can create applications accessing unique features of the platforms, distribute applications digitally to a broad audience and regularly update existing applications.*" Cfr. https://www.sec.gov/Archives/edgar/data/1439404/ 000119312511180285/ds1.htm, pag. 64.

[77] T. BRESNAHAN, J.P. DAVIS, PAI-LING YIN, (2014): *Economic Value Creation in Mobile Applications* in: The Changing Frontier: Rethinking Science and Innovation Policy, pp 233-286, National Bureau of Economic Research, Inc.

[78] Fonte: https://www.appannie.com/en/insights/market-data/app-economy-forecast-6-trillion-market-making/ *AppAnnie.*

[79] Fonte: https://www.appannie.com/en/insights/market-data/apps-used-2017/

[80] Ad esempio, sui dispositivi della *Apple* risulta pre-installata l'APP *Watch APP* utile, però, solo in abbinamento con lo specifico orologio prodotto dalla *Apple* (*iWatch*).

dispositivo perché preimpostate come applicazioni di sistema.[81] Le decisioni circa la pre-installazione delle APP vengono prese dai produttori di *handset* e/o dai fornitori dei sistemi operativi.

La maggioranza delle applicazioni si trova in veri e propri negozi virtuali denominati APP *store*: si tratta di piattaforme di distribuzione attraverso cui vengono rese disponibili al pubblico le APP. Una volta entrato nel negozio virtuale, rispetto al caso delle APP pre-installate, è l'utente che sceglie quale applicativo scaricare sul proprio *device* ed è quindi possibile, nonostante le politiche restrittive adottate dalle piattaforme, l'installazione di prodotti che potrebbero causare problemi di vulnerabilità dei sistemi operativi con il conseguente danneggiamento dell'apparecchio.[82]

Per motivi legati ai fenomeni di integrazione verticale, i negozi virtuali possono presentare delle incompatibilità con i sistemi operativi: ad esempio, l'APP *store* di *Google* funziona su *Android* ma non su terminali *iOS* e, viceversa, il negozio di *Apple* opera sul proprio sistema operativo ma non su *Android*. Tale aspetto è particolarmente importante perché idoneo a produrre effetti sul mercato e sui relativi assetti concorrenziali: in particolare, si rafforza, di fatto, la posizione di mercato dei principali *player* integrati verticalmente (in particolare *Google* e Apple). Inoltre, la posizione detenuta da questi ultimi sia nei sistemi operativi sia negli APP *store* costituisce, di fatto, una situazione di grande vantaggio nel mercato dei *big data*, in particolare nella fase della loro raccolta.

La diponibilità di applicativi nei negozi virtuali rientra tra le variabili che guidano la scelta degli utenti sul tipo di *device*, e connesso sistema operativo, da acquistare, innescando un vero e proprio **effetto di retroazione** (cd. esternalità di rete dirette e indirette) laddove più imprese sviluppano applicativi per un sistema operativo e il relativo negozio virtuale, maggiore sarà il numero di utenti che utilizzeranno l'APP *store* stesso, e ciò a sua volta condurrà a un maggior utilizzo del negozio da parte degli sviluppatori, e così via.

La varietà e la disponibilità di APP in uno *store* sono cresciute in maniera esponenziale dal 2008 ad oggi: a marzo 2017, sull'APP *store* di Apple erano disponibili 2,2 milioni di applicativi, mentre su *Google Play* 2,8 milioni. Un numero nettamente inferiore di APP è presente negli altri negozi virtuali. Occorre ricordare che tali "numeri" vengono superati velocemente, in ragione del fatto che ogni giorno gli APP *store* contengono nuove applicazioni.[83]

Dal punto di vista degli agenti coinvolti, sia dal lato della domanda (gli utenti finali) sia da quello dell'offerta (gli sviluppatori di APP), ai fini del presente Rapporto, appare quanto mai interessante un'analisi delle quote di mercato in termini di volumi sviluppati dagli *store*. Se, dal punto di vista dei ricavi, l'*Apple store* genera un fatturato ancora superiore a *Google Play*,[84] dal punto di vista dei volumi la situazione è esattamente opposta (**Figura 2.2**): lo *store* di *Google*, essendo inserito in un sistema aperto (v. *infra*), è

[81] Con la versione del sistema operativo *iOS 10,* ad esempio, la *Apple* consente di cancellare 23 APP pre-installate; in realtà non si tratta proprio di una vera cancellazione in quanto non libereranno memoria, anche se la *Apple* tiene a precisare che si tratta di applicazioni che nel complesso occupano solo 150 Mb di memoria. Per coloro che, dopo aver disinstallato una APP la vorrebbero riavere disponibile sul proprio apparecchio, basta semplicemente ritornare nel negozio virtuale della *Apple* e scaricare la APP. https://support.apple.com/en-gb/HT204221

[82] Attualmente sono disponibili all'utente numerosi negozi virtuali di APP su diverse piattaforme incluse quelle di *Microsoft*, *Google*, *Apple* e *Amazon*. È interessante osservare come nel tempo tali negozi virtuali abbiano notevolmente ampliato la gamma di prodotti offerti, dal momento che non vi si trovano solo APP ma anche contenuti (film, video, libri, ecc.)

[83] Secondo uno studio condotto da *pocketgamer.biz* in media in un giorno all'*Apple store* vengono sottoposte circa 1.000 APP e nel 90% dei casi l'approvazione da parte di *Apple* arriva in 7 giorni. http://www.pocketgamer.biz/metrics/app-store/.

[84] Secondo le analisi sviluppate dalla piattaforma *Statista.com*, ed in considerazione della scarsa rilevanza degli *store* alternativi all'*Apple store* e a *Google Play*, il primo detiene circa il 60% dei profitti congiunti, il secondo il 40% e, soprattutto, tale ripartizione sia prevista anche per i prossimi anni. https://www.statista.com/statistics/259510/revenue-distribution-between-the-apple-app-store-and-google-play/

leader in termini di numero di APP disponibili e caricate dagli sviluppatori e di *download* effettuati dagli utenti finali.

Il segmento degli APP *store*, quindi, appare caratterizzato da un elevato livello di concentrazione, con due operatori (*Google* e *Apple*) in una posizione di potere di mercato, anche in virtù dell'integrazione verticale e della quota detenuta nel collegato mercato dei sistemi operativi su *device* mobili.

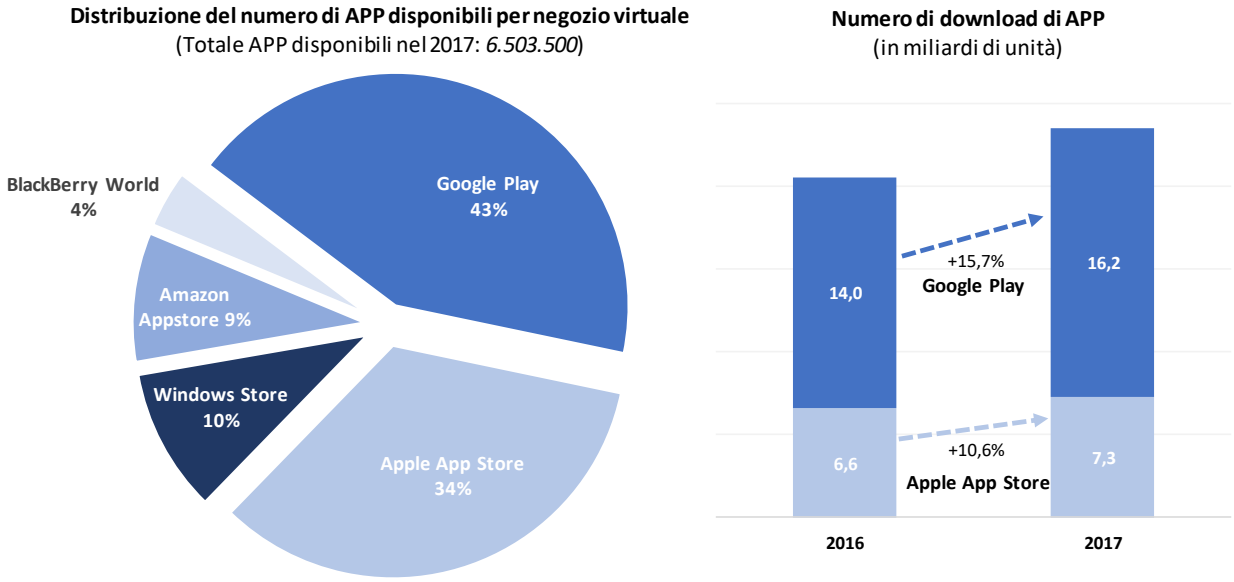


Figura 2.2 – Quote del mercato in volume (2017)
Fonte: Elaborazioni AGCOM su dati *Statista.com*

Sempre con riferimento alle caratteristiche strutturali, bisogna considerare che si tratta di un mercato in cui operano intense **esternalità di rete**; per il singolo utente, il beneficio derivante dall'acquisto e dall'uso di un bene o servizio aumenta al crescere del numero di altri utenti che utilizzano lo stesso bene o servizio (esternalità dirette), oppure di beni e servizi compatibili (esternalità indirette). Il maggior beneficio per il singolo utente (utilità) aumenta la sua disponibilità a pagare, incidendo in maniera notevole sul funzionamento del mercato, conducendo in particolare a un maggior grado di concentrazione.

L'ecosistema delle APP, inoltre, può essere inquadrato, dal punto di vista della teoria economica, nell'ambito dei mercati a due versanti (*two sided market*) o a più versanti (*multi sided market*), dato che gli attori interessati risultano più di due.[85] Come conseguenza, si manifestano anche **esternalità di rete incrociate**, quando le decisioni prese dagli individui appartenenti a un lato del mercato (gli utenti) producono degli effetti sugli agenti appartenenti all'altro lato (sviluppatori) e viceversa; nel caso delle APP, entrambi i lati del mercato traggono un beneficio dal numero di transazioni (*download*) effettuate tramite l'APP *store*. L'interesse degli utenti verso una specifica piattaforma, infatti, aumenta al crescere del numero di sviluppatori che vi operano e, viceversa, gli sviluppatori saranno portati a produrre APP per uno *store* all'aumentare degli utenti che vi accedono. L'intensità delle elasticità incrociate nelle due

[85] L'AGCOM ha ampiamente trattato l'argomento dei mercati a due versanti in precedenti indagini conoscitive, tra le quali è utile ricordare l'*Indagine Conoscitiva sulla Raccolta Pubblicitaria, Capitolo 1* (Allegato A alla delibera n. 551/12/CONS) e l'*Indagine Conoscitiva sul settore dei Servizi* internet *e sulla Pubblicità Online*, *Capitolo 1* (Allegato A alla delibera n. 19/14/CONS) e per ultimo nell'*Indagine Conoscitiva concernente lo sviluppo delle piattaforme digitali e dei servizi di comunicazione elettronica, Parte II – Piattaforme Digitali*. Per una rassegna della letteratura sui mercati a due versanti si veda M. RYSMAN, (2009): *The Economics of Two-Sided Markets*, Journal of Economic Perspectives 23. Nel caso delle APP uno *store* funge da intermediario tra gli sviluppatori di applicativi e gli utenti che usufruiscono dei servizi.

direzioni diventa un fattore rilevante nelle scelte strategiche relative al prezzo delle APP fissato dalle imprese.

In molti casi, in effetti, il consumo su un lato del mercato viene sussidiato dall'altro versante, attraverso la fissazione di prezzi bassi, spesso anche al di sotto del costo di produzione. Nel caso delle APP, ad esempio, un numero notevole di esse è disponibile agli utenti a un prezzo nullo, anche a fronte ovviamente di un costo di produzione positivo.[86]

Il manifestarsi delle esternalità di rete passa principalmente attraverso il conseguimento di quella che viene definita "massa critica di utenti": raggiunta tale quantità, infatti, la successiva crescita della rete segue un processo che si autoalimenta (effetto valanga o *bandwagon effect*). Il raggiungimento della massa critica può in parte spiegare perché le imprese, strategicamente, in una prima fase possono ritenere utile offrire il proprio prodotto ad un prezzo basso o, come spesso accade, gratuitamente; il prezzo basso infatti consente di attirare più rapidamente utenti e quindi di raggiungere prima la soglia critica, rispetto ad una politica di prezzo alto che in molti casi potrebbe scoraggiare l'acquisto del bene o servizio da parte dei consumatori.

La letteratura economica attribuisce alla presenza di esternalità di rete alcuni importanti effetti in termini di equilibrio di mercato; un primo effetto, di carattere generale, è quello per cui nei mercati con presenza di esternalità di rete si afferma un numero limitato di "standard tecnologici" (se non in molti casi un unico *standard* che opera in regime monopolistico), siano essi sistemi operativi, APP *store*, o singole categorie di APP (quali quelle di *social networking*).[87] Un ulteriore importante effetto si rinviene nel fatto che non è sempre la migliore tecnologia disponibile a prevalere sul mercato; proprio nella fase in cui si tende a raggiungere la soglia della massa critica, e quindi il punto di partenze dell' "effetto valanga", diventa cruciale la spinta che i consumatori iniziali danno alla diffusione del tipo di tecnologia e, di conseguenza, la tecnologia che prevarrà è fortemente influenzata dalle scelte iniziali di un certo numero di consumatori (effetto *lock-in*).

Tutti questi argomenti peculiari ai mercati a due (o più) versanti possono portare alla formazione di equilibri di mercato di tipo monopolistico o di oligopoli ristretti (*the winner takes all*), dove il vantaggio di chi supera per primo una massa critica di utenti (non necessariamente il primo operatore a fare ingresso sul mercato) diventa sempre più rilevante nel tempo, controbilanciando gli iniziali bassi costi

[86] In generale, quando le intensità degli effetti di rete non sono differenti nelle due direzioni, allora la possibilità di sussidiare un versante del mercato non è del tutto agevole. Al contrario, al crescere dell'asimmetria nelle intensità, può risultare utile sussidiare l'accesso a quel versante del mercato che genera più valore, determinando una situazione in cui aumenta il benessere dei consumatori. Tuttavia, la portata di tali effetti, dipende dal maggiore o minore livello di concorrenza in cui si trova ad operare una piattaforma, dal momento che la concorrenza impone dei vincoli che non consentono all'impresa la fissazione, per lei ottimale, dei prezzi come in una configurazione di mercato di tipo monopolistica. La concorrenza, infatti, porterebbe i prezzi a livello dei costi su entrambi i lati del mercato, non consentendo di praticare forme di sussidio.

[87] L'effetto delle esternalità di rete dipende anche da come vengono superati alcuni problemi legati da un lato alle scelte dei consumatori, dall'altro alle scelte delle imprese. Per quanto riguarda il lato della domanda, va sottolineato che il consumo di beni e servizi soggetti a esternalità di rete dipende dalla dimensione della rete stessa e soprattutto dalle aspettative degli utenti circa la dimensione che potrà raggiungere la rete; tale considerazione consente di affermare che saranno gli utenti stessi, in particolare quelli iniziali, a indirizzare la tecnologia che sarà più diffusa presso gli altri utenti (interdipendenza nelle scelte dei consumatori) tenendo presente la necessità di superare problemi di coordinamento nelle scelte conseguenti all'inerzia dei consumatori ad adottare nuove tecnologie o, al contrario, alla loro eccessiva mobilità che li porta ha "provare" tecnologie differenti. Le modalità attraverso cui vengono risolti questi problemi di coordinamento, determina, tra i molteplici equilibri possibili, quello che effettivamente si realizzerà. Dal lato dell'offerta, invece, le problematiche riguardano le modalità attraverso cui la tecnologia predominante verrà scelta o promossa; ad esempio, in presenza di esternalità di rete, in molti casi sono i governi a promuovere determinati *standard*. Nel caso in cui le scelte sugli *standard* siano lasciate al libero gioco del mercato, le imprese possono decidere di non rendere compatibili i propri prodotti, riducendo così la rete stessa, oppure di rendere compatibile la propria tecnologia. La mancata compatibilità, tuttavia, può essere una scelta voluta per innescare processi di *lock-in* (o cattura del consumatore) tali per cui per gli utenti i costi di transazione necessari per cambiare tecnologia (*switching cost*) diventano talmente alti da non agevolarne lo spostamento.

di ingresso sul mercato ed elevando le barriere all'affermazione di nuovi operatori. In tal senso, le APP sviluppate dai giganti tecnologici si sono evolute, rappresentando sempre più delle vere e proprie piattaforme attraverso cui veicolare contenuti agli utenti (*content consumption platform*). La APP di *Facebook*, in tal senso, rappresenta un caso emblematico (cfr. Capitolo 3); forte dell'enorme quantità e varietà di dati rilasciati dagli utenti e raccolti ed archiviati ad una velocità sempre più elevata, *Facebook* ha introdotto nel tempo numerose innovazioni nella fruizione di contenuti, in particolare di quelli video, e di accesso ad altri servizi (come ad esempio le APP che riguardano i giochi) in modo da rendere l'ambiente della sua APP sempre più simile ad una vera e propria piattaforma distributiva.

Inoltre, in considerazione del tempo crescente che gli utenti dedicano all'utilizzo dei *device* mobili, in particolare all'utilizzo di servizi online, l'ecosistema delle APP svolge un ruolo cruciale nell'acquisire risorse pubblicitarie come mostra la **Figura 2.3**. Per le caratteristiche del mercato poc'anzi descritte (presenza di economie ed esternalità di rete), di tale aumento hanno beneficiato soprattutto i giganti tecnologici come *Google* e *Facebook*.[88]

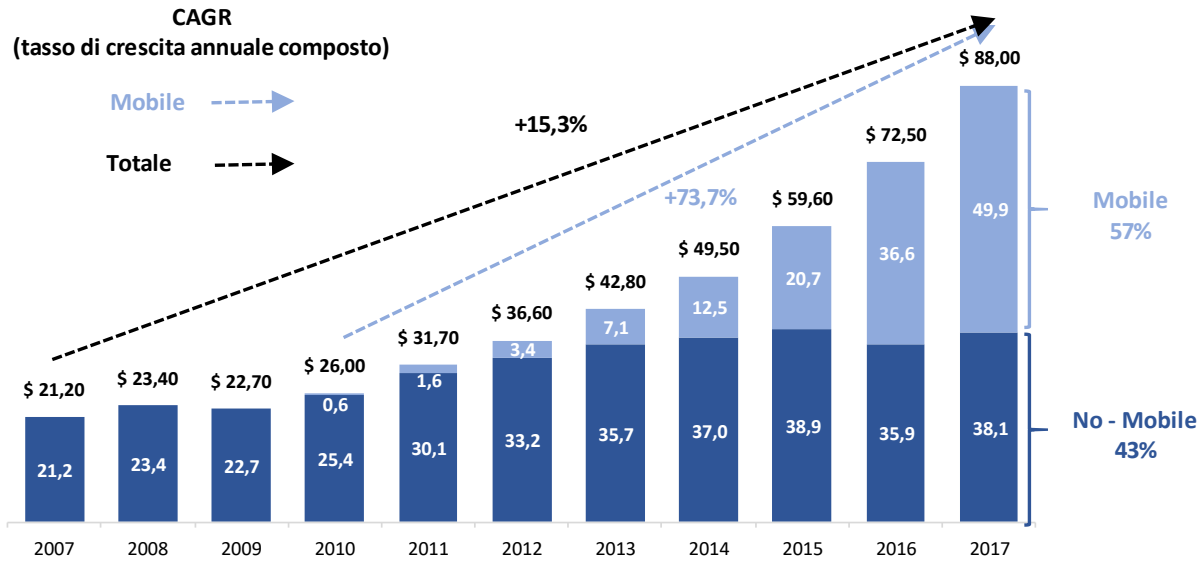


Figura 2.3 – Andamento dei ricavi pubblicitari online nel mondo (2007 - 2017)
Fonte: Elaborazioni AGCOM su dati *IAB*

L'ecosistema delle APP non appare di semplice definizione; oltre ad aspetti strettamente connessi al settore delle comunicazioni, infatti, bisogna considerare le influenze esercitate da numerosi altri settori industriali, tra i quali quello dell'intrattenimento, dei produttori di *device*, di hardware e di software. Il risultato è quello di avere un ecosistema assai complesso in cui la diversità appare essere un elemento rilevante; a seconda del posizionamento lungo la catena del valore, infatti, chi opera nell'ecosistema può adottare un modello differente di *business*, così come può andare incontro ad un diverso livello concorrenziale, al punto che alcune imprese sono allo stesso tempo in competizione e in collaborazione tra loro.

[88] Secondo i dati prodotti da *eMarketer*, in riferimento al mercato della pubblicità online americano, la quota di mercato congiunta di *Google* e *Facebook* nel 2018 era pari al 56,7% (37,2% per *Google* e 19,6% per *Facebook*). In termini previsionali, per gli anni futuri si prevede una leggera riduzione della quota di *Google* a vantaggio di un altro gigante tecnologico rappresentano da *Amazon*.
https://www.recode.net/2018/3/19/17139184/google-facebooks-share-digital-advertising-ad-market-could-decline-amazon-snapchat

Nell'ecosistema delle APP un ruolo primario è, come detto, svolto dagli *store*, che rappresentano delle piattaforme che fanno incontrare la domanda degli utenti di servizi internet in mobilità forniti attraverso applicazioni con l'offerta degli sviluppatori. Gli *store* si inseriscono in un contesto più complessivo di mercati integrati (legati tra loro attraverso relazioni di complementarità), che vanno dai sistemi operativi, ai produttori di *device* mobili, fino ai fornitori di servizi di telecomunicazioni. Questo insieme di servizi, prodotti e mercati forma una cd. piattaforma mobile.

Gli APP *store* sono negozi virtuali in cui gli utenti trovano gli applicativi per lo specifico sistema operativo installato sul *device*. Come ricordato in precedenza, il primo APP *store* di successo è stato quello realizzato da *Apple* nel 2008. Anche se in precedenza vi furono alcuni tentativi di creare negozi virtuali per lo scambio di software, è grazie all'idea e al servizio offerto da *Apple* che si è sviluppato il mercato.[89] È di quel periodo, infatti, il lancio del primo *iPhone*, uno strumento che ha modificato profondamente le abitudini di consumo di internet, innescando quel processo evolutivo che ha spinto sempre più verso la connettività in mobilità.

Allo *store* di *Apple* ha fatto seguito l'ingresso di *Google*, che presenta caratteristiche distintive. Infatti, *Apple* ha adottato un modello chiuso, con un maggior margine di guadagno, mentre *Google Play* risulta essere il negozio virtuale da cui è stato scaricato il maggior numero di APP e in cui viene caricato e reso disponibile agli utenti il maggior numero di applicazioni (*leadership* in volumi; v. **Figura 2.2**).

La valutazione dei consumatori riguardo una piattaforma mobile dipende molto da cosa essi trovano negli "scaffali" dei negozi virtuali, in particolare il numero di APP scaricabili e la loro varietà (le categorie). Allo stesso tempo, l'altro versante del mercato, quello degli sviluppatori, sarà attratto da *store* in cui è possibile raggiungere il più elevato numero possibile di potenziali clienti. Ciò innesca un fenomeno di *feedback* positivi tra i due versanti del mercato, che conduce, come sopra ampiamente esposto, a un elevato livello di concentrazione.

La **Figura 2.4** mostra la rapida crescita del numero di APP scaricate dagli utenti nel mondo su tutti gli APP *store* esistenti: in soli 8 anni, il numero di *download* è aumentato più di 10.000 volte rispetto al valore del 2009, mentre nell'ultimo anno si è registrata una crescita del 19,5%.

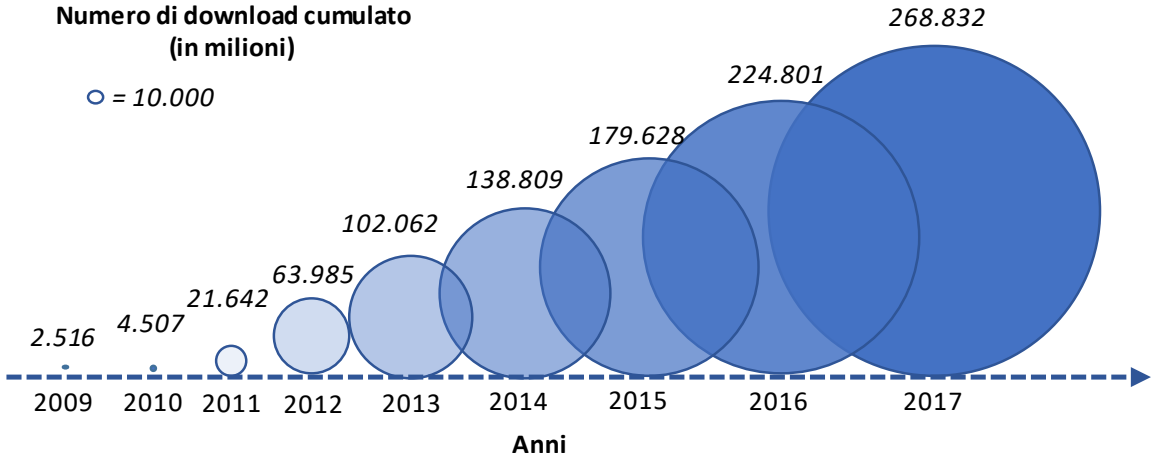


Figura 2.4 – Numero di applicativi mobili scaricati nel mondo dal 2009 al 2017 (in milioni)
Fonte: Elaborazioni AGCOM su dati annuali *Statista.com*

[89] J. WEST e M. MACE, (2010), *Browsing as the killer app: Explaining the rapid success of Apple iPhone*, Telecommunications Policy, 34 (5-6).

Secondo il sito specializzato nel settore *AppAnnie*, l'utilizzatore medio di smartphone possiede sul proprio *device* 80 APP (comprensive delle numerose APP preinstallate) e ne utilizza in media al mese 40.[90]

Con la diffusione degli smartphone e con l'aumento del numero di *download*, anche gli *store* devono adeguare la propria offerta, non solo in termini numerici ma anche di varietà dei prodotti e servizi offerti. Inoltre, quando i mercati a due versanti sono di successo, come quello delle APP, si genera un numero notevole di transazioni, tali per cui all'attore che funge da intermediario può ritornare utile fissare una commissione su ciascuna transazione. Al riguardo, si evidenzia che i due principali *store* prevedono, tra le disposizioni che gli sviluppatori devono accettare per poter fruire dei servizi della piattaforma, una commissione analoga e pari al 30% su ogni transazione effettuata.[91]

Essendo le APP scritte per specifici sistemi operativi e dal momento che quelli più diffusi al mondo sono due, non sorprende che i due *store* più grandi facciano capo proprio ai due sistemi operativi più diffusi (integrazione verticale: v. anche *supra*): *iOS* (*Apple's App Store*) e *Android* (*Google Play*).

Il negozio di APP *Google Play* è in assoluto quello più grande in termini di disponibilità di APP: nel 2017, erano disponibili oltre 2,8 milioni applicazioni, il 94% circa consente il *download* gratuito e forme di *in-APP purchase* (cioè vendite aggiuntive dopo il *download*), mentre il 70% circa risulta totalmente gratuito. Ciò ha generato ricavi complessivi per 9,8 miliardi di dollari.[92]

Nel negozio del sistema operativo *iOS*, a fine 2017, erano disponibili oltre 2,2 milioni di APP, mentre nel 2013 ne erano disponibili solo 300 mila. Circa 1 milione di APP sono definite "native" nel senso che sono sviluppate per funzionare appositamente con i *device* prodotti da *Apple*. Il giro d'affari a fine 2017 ha generato profitti per 11,4 miliardi di dollari (circa il 5% del fatturato totale della *Apple* rispetto all'1,5% del 2016); si pensi che in soli 7 giorni, dalla vigilia di Natale ad inizio anno 2018, si stima siano stati spesi nello *store* ben 890 milioni di dollari.[93]

Nonostante un numero così considerevole di applicazioni, se si fa riferimento a quelle più scaricate, risulta evidente quanto il mercato sia concentrato nelle mani di poche imprese, per giunta quelle più importanti nell'ambito dell'economia digitale, come *Google* e *Facebook* (**Figura 2.5**). Ciò evidenzia nuovamente un elevato grado di integrazione dei soggetti lungo tutta la filiera dei servizi online.

[90] https://www.appannie.com/en/insights/market-data/app-annie-2017-retrospective/

[91] Per *Google* Play si veda https://support.google.com/googleplay/android-developer/answer/112622?hl=it, che prevede: *"Per le applicazioni e i prodotti in-app che vendi su Google Play, la commissione sulle transazioni è pari al 30% del prezzo. Ricevi il 70% del pagamento. Il restante 30% è destinato al partner di distribuzione e alle commissioni dell'operazione.* Per l'*APP store* di Apple si veda https://developer.apple.com/programs/whats-included/ , dove si indica testualmente: "*You get 70% of sales revenue. 85% for qualifying subscriptions*"

[92] Fonte: *statista.com* https://www.statista.com/statistics/444476/google-play-annual-revenue/

[93] Fonte: *forbes.com* https://www.forbes.com/sites/chuckjones/2018/01/06/apples-app-store-generated-over-11-billion-in-revenue-for-the-company-last-year/#7ceb836f6613

Figura 2.5 – Top 10 APP per *download* nei due principali *store* (2017)
Fonte: Elaborazioni AGCOM su dati *AppAnnie.com* e *AppBrain.com*

Al riguardo, gli effetti derivanti dall'integrazione verticale con i sistemi operativi non risultano evidenti dai dati prodotti in **Figura 2.5** giacché, come detto, la maggioranza delle APP sviluppate dagli operatori integrati (*Google* e *Apple)*, risulta pre-installata sui *device*. Secondo le statistiche prodotte da *appbrain.com*, infatti, l'APP *Google Play services*, preinstallata sui *device* con sistema operativo *Android* e che consente l'accesso al negozio virtuale, risulta aver avuto più di 5 miliardi di installazioni dalla sua comparsa ad oggi.

Ciò detto, le APP con elevata diffusione realizzate da sviluppatori "indipendenti", ossia non integrati in nessun altro stadio della catena del valore, sono assai rare (nel caso dell'*Apple store*, *Twitter* risulta l'unica tra le prime dieci; lato destro della **Figura 2.5**). Tale evidenza fa emergere **l'importanza dell'integrazione nei vari stadi della filiera ai fini della posizione acquisita dagli operatori in ciascuno di essi. Evidentemente, questo effetto si riverbera anche sui processi di acquisizione dei dati, e sugli assetti competitivi in tali ambiti**.

Al riguardo, è utile ricordare che circa il 94% degli applicativi risulta *free*, vale a dire è reso disponibile per gli utenti in maniera gratuita.[94] Tra l'altro, in molti casi, si tratta di un modello di monetizzazione che è più corretto definire *freemium*; ossia dalla combinazione dei termini *free* (gratuito – non a pagamento) con *premium* (sovrapprezzo). Ciò comporta la cessione gratuita della parte principale (cd. *core*) del prodotto, mentre prevede la vendita dietro corrispettivo di prodotti addizionali (definiti *premium*). Tale *strategia di prezzo* non solo è mirata a sfruttare l'effetto positivo di retroazione (*feedback*) derivante dalle esternalità di rete (v. *supra*), ma è anche **volta all'acquisizione del maggior numero di dati degli utenti, che vengono poi monetizzati dagli operatori, direttamente e/o indirettamente (ossia cedendoli a terzi), attraverso vari usi.**

La **Figura 2.6** mostra la distribuzione delle APP dei due principali negozi virtuali (*Apple Store* e *Google Play*) per tipologia di categoria. Ferma restando la non perfetta corrispondenza tra le categorie, dal momento che ciascuna piattaforma classifica le APP secondo una propria metodologia, emerge una distribuzione molto simile. In effetti, bisogna ricordare che numerose APP (specie tra le più diffuse) sono

[94] Al 19 febbraio 2017, su *GooglePlay* erano disponibili 2.734.073 APP, di cui 2.520.462 gratuite, pari al 92%, e 213.611 a pagamento, pari all'8% (fonte: *Appbrain.com*).

sviluppate nella duplice versione al fine di essere vendibili in entrambi i negozi virtuali e raggiungere così una più ampia platea di potenziali utenti.

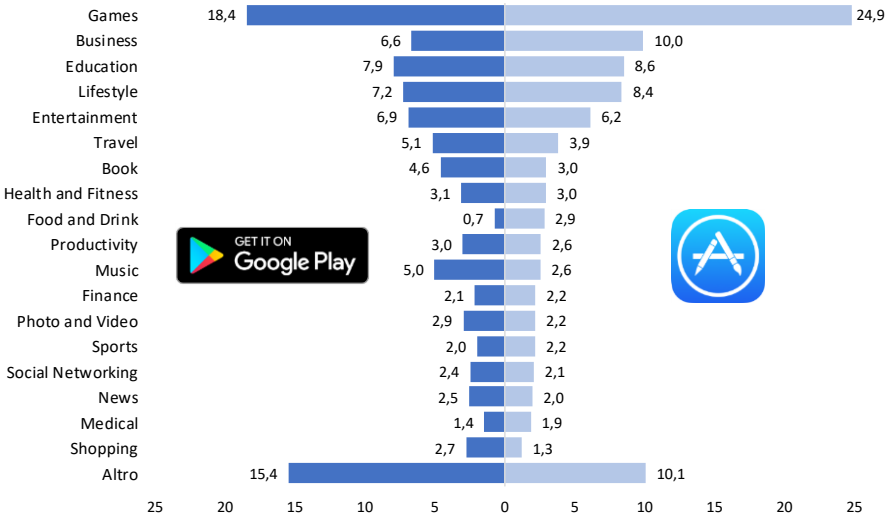


Figura 2.6 – APP per principali categorie nel 2017 (%)
Fonte: Elaborazioni AGCOM su dati *Statista.com* e *AppBrain.com*

Emerge, in primo luogo, il rilevante ruolo degli applicativi sviluppati per i giochi; tale categoria è quella che, oltre a coinvolgere l'utente per più tempo, produce il maggior flusso di entrate dirette (ossia di ricavi generati direttamente tramite la vendita del servizio agli utenti finali). A titolo di esempio, secondo *Statista.com*, nel 2017 i giochi *Arena of Valor* e *Fantasy Westward Journey* hanno generato più di 3,4 milioni di dollari di ricavi nel 2017.[95]

In secondo luogo, è interessante notare come le categorie *lifestyle* e *entertainment* siano quelle che raggiungono la maggior parte degli utenti, essendo per loro natura di APP trasversali: si pensi che circa il 78% dei consumatori che posseggono un sistema operativo *iOS* ha scaricato almeno un'APP appartenente a una di queste due categorie.

Dato che, come illustrato in precedenza, i proprietari degli APP *store* svolgono un ruolo chiave nel processo di intermediazione tra sviluppatori ed utenti, questi sono in grado di guidare le scelte dei consumatori, orientandoli in vari modi verso le diverse applicazioni. Sono questi, infatti, che decidono le condizioni di accettazione, pubblicazione, diffusione delle APP nello *store*.

2.6. Una soluzione di mercato alle transazioni di dati: i permessi

Le dimensioni del mercato descritte in precedenza comportano un enorme e crescente flusso di dati la cui origine è l'individuo; le informazioni raccolte vengono utilizzate per migliorare i prodotti offerti e rendere, quindi, l'esperienza di consumo per l'utente più gradevole e, allo stesso tempo, per costituire banche dati da cui è possibile estrarre valore in numerosi modi (cd. usi primari e secondari). Questi dati possono essere raccolti direttamente tramite un'interfaccia utente, come una APP, ma anche indirettamente, vale a dire senza che vi sia una specifica attività da parte dell'utente, tramite i numeri di telefono che, dato il carattere personale del *device*, rappresentano dei veri e propri identificativi univoci (UDID – *Unique Device Identifier*).

[95] Fonte: https://www.statista.com/statistics/505625/leading-mobile-games-by-global-revenue/

Le informazioni che vengono raccolte, potenzialmente, sono tutte quelle disponibili su un *device* mobile e relative a tutte le attività svolte dall'utente, quali il salvataggio di dati, la lista dei contatti, i video, le foto, i messaggi, la posta elettronica e le varie *password* utilizzate per accedere a specifici servizi, come quelli finanziari. Un ruolo sempre maggiore è assunto dalle informazioni concernenti la localizzazione dell'utente attraverso cui è possibile raggiungere l'utente in qualsiasi luogo, fornire risposte adeguate alle proprie esigenze, come la ricerca di posti per pranzare o per il pernottamento.

Lo sviluppo tecnologico ha consentito la diffusione sempre maggiore dei processi di "datizzazione", ossia della trasformazione di qualsiasi cosa (film, libri, messaggi vocali, movimenti del corpo, ecc.) in formato digitale. A puro titolo di esempio, è possibile affermare che *Facebook* abbia datizzato le relazioni sociali, *LinkedIn* quelle lavorative e *Twitter* le opinioni, ma gli esempi potrebbero riguardare anche le APP che offrono servizi di *health*.

Il grado di consapevolezza degli utenti in merito alla cessione delle proprie informazioni è un aspetto centrale: secondo una ricerca condotta dall'Unione Europea nel 2015 (*Special Eurobarometer 431 - Data Protection*), il 69% degli utenti italiani ritiene la cessione delle informazioni individuali una conseguenza naturale degli attuali stili di vita, e il 54% la mette in connessione con l'accesso ai servizi digitali. Al contempo, il 52% sottolinea come la tendenza a fornire dati individuali non sia priva di rischi.

Tra i rischi maggiormente percepiti, vi sono l'utilizzo da parte delle *internet company*, direttamente o tramite la cessione a terzi, delle proprie informazioni senza il consenso e la possibilità di essere vittime di frodi, di subire furti di identità e di ricevere pubblicità indesiderata. Conseguentemente, i cittadini si aspettano regole che consentano una migliore protezione delle informazioni private; in particolare, quando si verificano episodi di mal utilizzo delle informazioni (*breach*), gli utenti, nel 51% dei casi, auspicano che all'impresa non debba essere più consentito l'utilizzo dei dati individuali e nel 39% dei casi un risarcimento del danno.

D'altra parte, è convinzione dell'Autorità che, laddove possibile, le analisi di tematiche inerenti al mondo dei *big data* debbano essere più proficuamente svolte attraverso metodologie coerenti con l'ecosistema oggetto di osservazione. Pertanto, **lo studio successivo** (v. in particolare il paragrafo 2.7) è stato condotto su **milioni di osservazioni riguardanti comportamenti in rete effettivamente adottati dagli utenti** (e non attraverso *survey* realizzate, per mezzo di risposte a questionari, su limitati campioni di cittadini).

In particolare, l'analisi si focalizza sui negozi virtuali. Infatti, gli APP *store* sono un importante esempio di servizio/modalità attraverso cui vengono scambiati dati digitali per mezzo delle piattaforme online. Le modalità di trattamento dei dati vengono rese note agli utenti prima del download di un'APP, tramite un accordo di licenza (EULA – *End-User License Agreement*). In alcuni casi, per il corretto funzionamento di un'APP è strettamente necessario consentire l'accesso ad alcune componenti hardware e software dello smartphone e ad alcune informazioni con un certo grado di sensibilità: ad esempio, se si è interessati ad un'applicazione che riporta le previsioni meteorologiche, è necessario che essa possa accedere alle informazioni relative alla localizzazione dell'utente, al fine di consentire alla APP di fornire le informazioni meteo sul luogo in cui si trova in quel momento l'utente.

Per quanto sopra esposto, appare quanto mai naturale analizzare gli APP *store*, e in particolare quello di *Google*, come esempio di esito di mercato alla gestione dei dati online. Infatti, quanto illustrato in precedenza mostra con forza:

a) la crescente, e oramai maggioritaria tra gli utenti, connettività da dispositivi mobili (v. **Figura 2.1**);

b) l'uso delle piattaforme mobili (*device*, sistema operativo, APP *store*) come strumenti personali non solo di comunicazione ma anche di gestione di una serie di attività quotidiane svolte dai cittadini (cfr. paragrafo 2.2);
c) l'affermazione delle APP, e quindi dei negozi virtuali da cui possono essere scaricate, come intermediari che facilitano lo svolgimento di queste attività (v. **Figura 2.4** e **Figura 2.6**);
d) l'affermazione di *Google Play* come *store* privilegiato sia per l'attività di download delle APP degli utenti finali, sia per quella di distribuzione degli applicativi mobili da parte degli sviluppatori (v. **Figura 2.2**).

In questo contesto, gli APP *store* si sono dotati di specifiche politiche di trattamento dei dati, al fine sia di rispettare le norme sulla protezione dei dati vigenti nei singoli Stati, sia di creare un rapporto di maggior fiducia con gli utenti dei propri servizi.

Privacy Permission è il termine utilizzato da *Google Play* per indicare il fatto che gli utenti, attraverso la concessione di permessi, cedano una serie di informazioni per poter installare e utilizzare un'APP e usufruire dei suoi servizi**. Il sistema dei permessi, quindi, rappresenta il meccanismo adottato da questa componente del mercato, per mezzo del quale viene disciplinata la cessione di dati dall'utente allo sviluppatore dell'APP**. Tale sistema viene definito e monitorato dal gestore dello *store*.

Gli sviluppatori, per poter distribuire i loro prodotti sullo *store* di *Google Play*, oltre ad utilizzare una serie di *tool* tecnici messi a disposizione sulla piattaforma appositamente creata da *Android*[96], devono sottoscrivere un contratto di distribuzione che prevede esplicitamente che "*Lo Sviluppatore accetta di proteggere la privacy e i diritti legali degli utenti se rende i suoi Prodotti disponibili tramite Google Play. Se gli utenti le forniscono, o il suo Prodotto accede a o utilizza, nomi utente, password o altri dati di accesso o informazioni personali, deve comunicare agli utenti che le informazioni saranno disponibili per il suo Prodotto, nonché fornire un'informativa sulla privacy legalmente adeguata e protezione a tali utenti*." (articolo 4.8, Contratto di distribuzione per gli sviluppatori *Google Play* aggiornato a febbraio 2018).[97]

Per essere scaricate dall'utente e funzionare correttamente, le applicazioni possono richiedere l'accesso sia ad alcune funzionalità dei *device*, sia ad alcune informazioni ivi contenute. Proprio a causa delle molteplici funzionalità di smartphone e *device* mobili, questi dispositivi producono e conservano una notevole quantità di informazioni e dati riferibili al singolo individuo, dalla geo-localizzazione al registro di chiamate e messaggi. Si tratta in molti casi di informazioni che possono essere cruciali per il funzionamento delle APP, ma che richiedono il consenso al relativo uso da parte degli utenti. Non a caso, quando si effettua il *download* dell'APP, il sistema operativo chiede all'utente di selezionare la casella "accetto" dopo aver indicato tutte le informazioni a cui l'APP dovrà accedere. I meccanismi attraverso cui gli sviluppatori rivelano come le proprie APP interagiscono con i *device* degli utenti e con le informazioni individuali veicolate dagli stessi dispositivi vengono, come detto, chiamati "*permessi*".

Ci sono molti contesti (*blog*, siti specializzati, la pagina di *Android*, ecc.) in cui gli utenti possono trovare informazioni di dettaglio sui permessi che un'APP richiede. Il caso più evidente è la schermata che appare sul *device* dell'utente nel momento in cui quest'ultimo sceglie di scaricare un'applicazione sul proprio dispositivo. Tipicamente, su uno smartphone (o un tablet) con sistema operativo *Android*, una volta che l'utente clicca sull'icona "*installa*", appare una schermata come quella riportata in **Figura 2.7**.

[96] https://developer.android.com/index.html.
[97] https://play.google.com/about/developer-distribution-agreement.html.

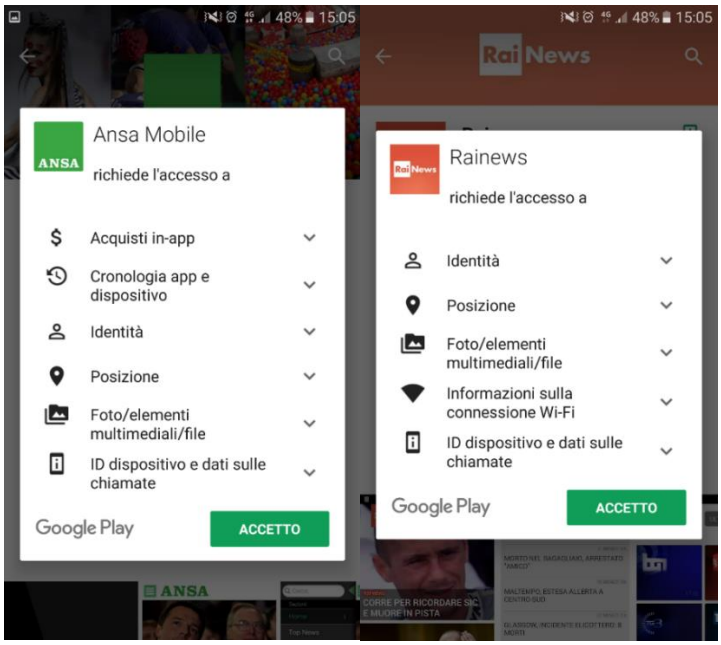


Figura 2.7 – *Screenshot* dei permessi richiesti da due *APP* di informazione
Fonte: *Google Play Store*

Inoltre, prima che l'APP venga installata, gli utenti possono verificare i permessi richiesti accedendo all'opzione "*dettagli autorizzazione*" di *Google Play*, come riportato in **Figura 2.8**. La lista dei permessi viene aggiornata ogni qualvolta l'APP stessa viene aggiornata.

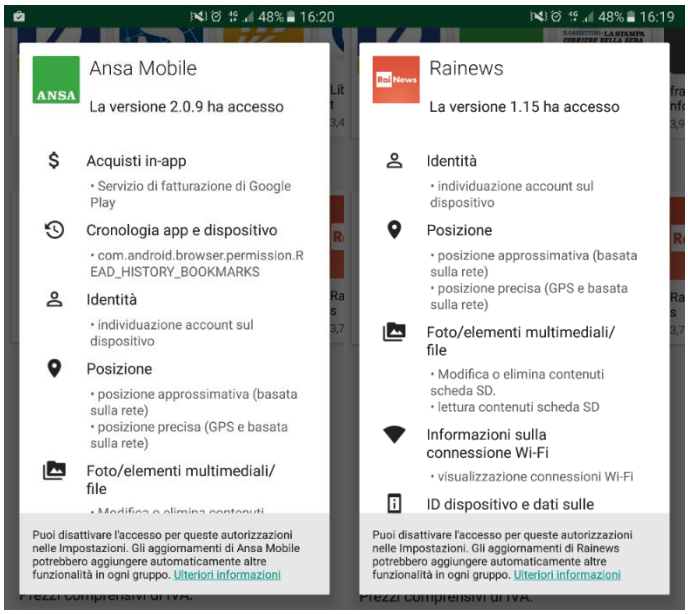


Figura 2.8 – *Screenshot* dei dettagli autorizzazione richiesti da due APP di informazione
Fonte: *Google Play Store*

I permessi possono consentire alla APP di accedere a numerosi dati che riguardano gli utenti, relativi, ad esempio, alle attività ricreative e agli spostamenti, alle abitudini di navigazione e di acquisto, al consumo dei media, alle foto e ai video scattati e condivisi.

Da una *survey* condotta dal *Pew Research Center* sugli utenti statunitensi emerge una forte consapevolezza sulle informazioni richieste dalle APP, visto che il 90% degli intervistati sottolinea che tali informazioni

sono (molto o abbastanza) importanti nella scelta di effettuare (o meno) il *download* di un'applicazione, e che il 60% degli stessi ha preferito non scaricare una APP dopo aver scoperto che le informazioni sottostanti ai permessi di accesso richiesti, in molti casi, non risultavano necessarie al buon funzionamento della APP.[98]

Nel sistema operativo *Android*, le APP richiedono l'adesione degli utenti alle condizioni di uso al momento della loro installazione e, contemporaneamente, gli utenti possono prendere visione di una breve descrizione del permesso. I metodi per informare l'utente su come i loro dati verranno utilizzati da un'APP sono un punto di contatto tra l'utente, *Google* (azienda fornitrice e sviluppatrice del sistema operativo e dello *store*) e gli sviluppatori terzi di applicazioni mobili.

I "permessi", come detto, sono la modalità attraverso cui *Google* chiede agli sviluppatori di rivelare come le loro APP interagiranno con i *device* dell'utente e di quali informazioni esse hanno bisogno (o che comunque vengono richieste). All'interno delle pagine dedicate agli sviluppatori su sistema *Android*, è possibile, inoltre, rinvenire una classificazione più specifica operata dalla società in base alle funzionalità a cui la APP richiede di accedere, così come riportato nella successiva **Tabella 2.1**.

Tabella 2.1: Categorie di permessi[99]

Categoria	Breve descrizione
Calendario	Permessi *run-time* relativi al calendario degli utenti
Camera	Permessi associati all'uso della camera o al salvataggio di fotografie o video dal *device*
Contatti	Permessi *run-time* relativi ai contatti e ai profili sul *device*
Posizione	Permessi che permettono di accedere alla posizione del *device*
Microfono	Permessi che permettono di accedere al microfono audio del *device*
Telefono	Permessi associati a caratteristiche e funzionalità del telefono
Sensori	Permessi associati all'uso della camera o al salvataggio di fotografie o video dal device
SMS	Permessi *run-time* relativi agli SMS degli utenti
Memoria	Permessi *run-time* relativi alla memoria esterna condivisa

Fonte: developer.android.com

Due attributi sono rilevanti nel sistema dei permessi di *Google Play*: l'appartenenza ad un gruppo di permessi (*Permission Group*) e il livello di protezione (*Protection Level*). Il primo attributo serve a raggruppare i permessi per poi presentarli all'utente durante il processo di installazione, mentre il *Protection Level* specifica come si dovrà comportare il sistema operativo al momento dell'installazione della APP e, quindi, se sarà necessario chiedere il consenso all'utente.

I gruppi di permessi, quindi, hanno la funzione principale di facilitare la comprensione al pubblico. Una simile architettura dovrebbe consentire una maggiore comprensibilità dei permessi, rendendo così le scelte più consapevoli nel momento in cui l'utente si trova a dover decidere se dare il proprio consenso.

[98] Pew Research Center (2015), *Apps Permissions in the Google Play Store*, www.pewinternet.org. La *survey* è stata condotta su un sotto-campione di 461 adulti (più di 18 anni) provenienti dal Knowledge Panel di GfK Group.
[99] https://developer.android.com/reference/android/Manifest.permission_group.html .

Ciascun gruppo di permessi può comprendere anche numerosi permessi singoli che rispondono ai criteri stabiliti dalla piattaforma *Android* (**Figura 2.9**).

Di conseguenza, quando la APP fa richiesta di accesso ad uno specifico permesso, l'utente finale dovrà concedere il permesso all'intera categoria. A titolo di esempio, si supponga che una APP abbia bisogno del permesso di monitorare, modificare o interrompere una chiamata in uscita (permesso denominato - *process outgoing calls*), come accade per le applicazioni VOIP (*Voice Over internet Protocol*); lo sviluppatore dovrà indicarlo al gestore della piattaforma, dichiarandolo nel *permission manifest*; all'utente tuttavia comparirà solo la richiesta di prestare consenso all'intero gruppo di permessi a cui lo stesso appartiene (ossia il gruppo "telefono").

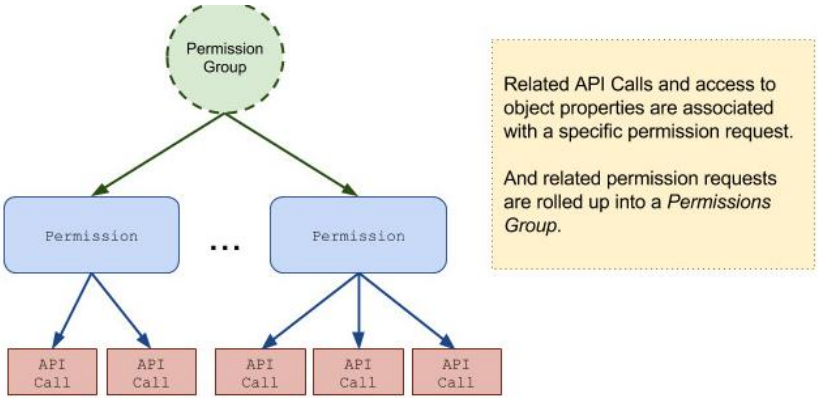


Figura 2.9 – Architettura dei permessi in Android
Fonte: developer.android.com

Alcuni permessi non hanno lo scopo principale di raccogliere informazioni, seppure lo fanno, quanto quello di garantire un'interazione tra la APP e il sistema operativo del *device* in modo da garantire il buon funzionamento dell'applicativo assicurando, al contempo, un livello minimo di sicurezza.

In tal senso, ***Android* opera una basilare distinzione in base ai livelli di protezione da attribuire ai singoli permessi, individuando tre livelli di protezione che gli sviluppatori devono considerare: *normal*, *dangerous* e *signature***. Tale distinzione non è fatta considerando i soli pericoli relativi al trattamento dei dati dell'utente, ma principalmente per tener conto del livello di rischio associato al corretto funzionamento dell'hardware.

I permessi con l'attributo *normal* sono quelli che presentano un basso rischio per le altre applicazioni, per il sistema operativo e per gli utenti. Il sistema automaticamente consente l'accesso a questa tipologia di permessi, senza neanche chiedere il consenso esplicito agli utenti i quali, però, possono sempre decidere di non installare la APP. I permessi con l'attributo *dangerous* presentano un rischio maggiore, dal momento che consentono l'accesso a dati dell'utente oppure richiedono il controllo sull'apparecchio. Poiché queste interrogazioni della APP sono potenzialmente rischiose, viene chiesto un consenso esplicito da parte dell'utente, o comunque la richiesta deve passare attraverso un'esplicita notifica all'utente. L'attributo *signature*, invece, fa riferimento a permessi la cui pericolosità è assimilabile a quella dei permessi *normal*, ma che riguardano applicazioni che presentano la stessa sigla (o certificato) usata per firmare l'applicazione che per prima ha certificato il permesso. Se i due certificati combaciano, allora il permesso viene concesso in automatico, senza il consenso esplicito dell'utente. Un livello aggiuntivo di protezione per i permessi *signature* è rappresentato dall'attributo *signatureOrsystem* che consente di agganciare un permesso con attributo *signature* anche alle altre applicazioni del sistema operativo; tale attributo viene spesso utilizzato dagli sviluppatori di applicazioni delle grandi imprese, in modo da garantire un vantaggio alle loro stesse applicazioni.

Occorre ricordare che per gli sviluppatori di APP è comunque possibile individuare nuovi specifici permessi; inoltre, i singoli permessi, i gruppi e i livelli di rischio definiti dal gestore dello *store* (in questo caso *Google*) sono comunque soggetti a continue revisioni in risposta alle funzionalità sempre più sofisticate dei *device*, alle esigenze di privacy degli utenti e, chiaramente, alla sicurezza del sistema operativo.

Ai fini del presente studio, per una più rigorosa individuazione dei permessi che consenta di superare la mera ottica tecnica-informatica, nel paragrafo successivo verranno esposte ulteriori e più fini classificazioni.

In conclusione, vale osservare come la gestione dei dati relativi ad un singolo individuo nel mondo digitale passi attraverso strumenti quali quello dei permessi (cd. *permission*); tramite questo sistema nessuna applicazione, per *default*, ha il permesso di eseguire una qualsiasi funzione che possa impattare in maniera negativa sul funzionamento delle altre applicazioni, sul sistema operativo e, chiaramente, sull'utente.

I permessi, quindi, disciplinano la parte dei rapporti che intercorrono tra utenti e sviluppatori di APP relativa al flusso di dati e il loro trattamento. Un'analisi di tali strumenti permette di capire il reale comportamento di consumatori e imprese nell'ambito di questo scambio, nonché di valutarne l'efficienza economica e sociale.

In questo quadro, è da notare, infine, come l'architettura dei permessi, anche sulla base delle normative nazionali in materia di privacy, venga definita e classificata dai **proprietari delle piattaforme mobili (ossia sistemi operativi, e APP *store*) che, quindi, si trovano in una posizione privilegiata e in grado di orientare il mercato dei dati.**

2.7. L'esistenza di uno scambio implicito tra utenti e operatori web

Alcune recenti ricerche hanno evidenziato la presenza di un nesso causale tra il numero di permessi richiesti da una APP e il comportamento degli utenti. L'OECD[100] rileva che la politica dei permessi è strategicamente rilevante per il successo di un'APP, dal momento che gli utenti si mostrerebbero attenti alla quantità e alla tipologia delle informazioni che vengono cedute nel momento in cui si decide di scaricare e usare una APP.

Tuttavia, gli studi empirici in materia mostrano una serie di risultati contrastanti riguardo alla valutazione che gli utenti danno alla raccolta dei dati da parte delle APP. Il lavoro di Grossklags & Acquisti (2007), ad esempio, fa emergere un esito non scontato tra la disponibilità massima a pagare degli utenti in cambio di una minore richiesta di dati (*willingness to pay to protect*) e la disponibilità minima a cedere i dati (*willingness to accept*); per gli autori, infatti, i processi sottostanti alle decisioni degli utenti sono differenti a seconda che si tratta di proteggere i dati o, viceversa, di fornire dati; la disponibilità a pagare per essere protetti risulta più bassa.[101]

Il lavoro di Savarge & Waldman (2014), sulla base di un campione di individui a cui è stata chiesta la disponibilità a pagare per poter rinunciare a cedere alcune informazioni, stima in 2,28$ la disponibilità a pagare *una tantum* per poter vedere "cancellati" i dati relativi alle ricerche effettuate sul web, e di 4,05$ per l'occultamento dei dati relativi alla propria agenda dei contatti.[102]

Kummer & Schulte (2016), con un'analisi simile a quella del presente lavoro, testano con successo l'ipotesi che gli sviluppatori offrano le APP a un prezzo più basso in cambio dell'ottenimento di una

[100] OECD, (2013); *The App economy*.
[101] J. GROSSKLAGS e A. ACQUISTI, (2007), *When 25 Cents is too much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information*, WEISS.
[102] S. J. SAVAGE e D. M. WALDMAN, (2014), *The Value of Online Privacy: Evidence from Smartphone Applications*, Technical Report.

quantità e/o qualità maggiore di dati digitali prodotti dagli utenti. Essi trovano che sia la domanda di APP, sia l'offerta sono influenzate in maniera significativa dal numero di permessi richiesti, facendo emergere il *trade off* esistente tra il lato della domanda e quello dell'offerta.[103]

Dall'analisi di questi lavori emerge, inoltre, che **le APP a pagamento presentano mediamente un numero di permessi inferiore rispetto a quelle gratuite**. **In maniera implicita, dunque, il mercato sembra attribuire un valore economico ai dati digitali dell'utente: un loro minore rilascio, infatti, presuppone un prezzo più alto da pagare per l'utente**.

2.7.1. Lo studio su (milioni di) APP e permessi

Nel presente lavoro, svolto dal Servizio economico-statistico dell'Autorità con la collaborazione del Dipartimento di Ingegneria Informatica Automatica e Gestionale dell'Università "la Sapienza" di Roma,[104] si proseguirà nel solco di questo filone di ricerca, sfruttando le informazioni ricavabili da un ricco *dataset* di APP e dei relativi permessi richiesti.

Il *dataset* comprende informazioni su 1.135.700 APP presenti su *Google Play*, vale a dire circa l'80% delle APP disponibili sullo *store*. Si tratta di informazioni raccolte sulla base di un processo chiamato *crawling*. Il rimanente 20% fa parte di una quota residuale di applicazioni, appartenenti alla "coda lunga", che risultavano solo marginalmente diffuse, in termini di *download*, presso gli utenti finali.

Per ciascuna APP, le informazioni raccolte riguardano:

- la categoria di appartenenza;
- il prezzo;
- la soglia minima dell'*in-app purchase*;
- la soglia massima dell'*in-app purchase*;
- il *rating* assegnato dagli utenti;
- il numero delle *review* scritte dagli utenti;
- il numero dei *download*;
- le tipologie di permessi richiesti.

Per quanto riguarda le **categorie**, esse rappresentano uno strumento fondamentale, dato che la relativa classificazione consente all'utente di navigare più facilmente all'interno di uno *store* che contiene circa 3 milioni di applicativi. Difatti, con la crescita esponenziale del numero delle APP, è diventato rilevante per il gestore della piattaforma introdurre scorciatoie che rendano intuibile e facile la navigazione degli utenti all'interno del negozio. A titolo di esempio, nel luglio 2016, sono state modificate alcune categorie esistenti e aggiunte delle altre.[105]

La **Tabella 2.2** mostra la distribuzione delle APP per categorie; è importante sottolineare che la categoria dei giochi contiene al suo interno 30 sottocategorie. I dati raccolti sono in linea con quanto già mostrato nella **Figura 2.6**.

---

[103] KUMMER M. E., SCHULTE P., (2016), *When private information settles the bill: money and privacy in Google's market for smartphone applications*, ECONSTOR.

[104] Il processo di acquisizione di dati e informazioni sulle APP dallo *store* di *Google* è stato svolto dal Dipartimento di Ingegneria Informatica Automatica e Gestionale dell'Università "La Sapienza" di Roma, con cui il Servizio economico-statistico dell'Autorità ha condiviso il progetto di ricerca. Si ringraziano, in particolare, il Prof. A. Vitaletti e l'Ing. A. De Carolis.

[105] Ad esempio, dal 27 luglio 2016, su *Google Play* sono previste 8 nuove categorie; Arte e Design, Auto e Veicoli, Bellezza, Incontri, Eventi, Mangiare e Bere, Casa e Arredamento e Genitori. Altre due categorie sono state rinominate; Trasporti in Mappe e Navigatori, mentre Media e Video in Strumenti Video. http://www.androidauthority.com/google-play-store-new-app-categories-706028/

Tabella 2.2: Distribuzione degli applicativi per categoria

Categoria	numero di APP	%
Giochi	228.823	20,2
Istruzione	100.293	8,83
Strumenti	82.799	7,29
Intrattenimento	80.915	7,12
Lifestyle	79.158	6,97
Personalizzazione	72.457	6,38
Affari	64.551	5,68
Libri e consultazione	59.990	5,28
Viaggi e info locali	49.093	4,32
Musica e audio	41.793	3,68
Produttività	33.821	2,98
Notizie e riviste	33.002	2,91
Medicina	32.844	2,89
Finanza	25.793	2,27
Comunicazione	25.462	2,24
Social	22.246	1,96
Shopping	19.039	1,68
Mappe e navigatori	17.618	1,55
Fotografia	16.859	1,48
Medicina	16.646	1,47
Strumenti video	14.843	1,31
Famiglia	6.313	0,56
Meteo	4.632	0,41
Fumetti	3.460	0,3
Librerie e demo	3.250	0,29
Totale	**1.135.700**	**100**

Fonte: Elaborazioni AGCOM su dati *Google Play*

Come descritto in precedenza, ai fini del corretto funzionamento, una APP deve richiedere di poter aver accesso a una serie di informazioni che riguardano sia la componente hardware del *device* sia i dati dell'utente. Quello dei permessi è, quindi, un sistema complesso, che ha al suo interno una serie di problematiche classificatorie. **Alcuni permessi svolgono una funzione tecnica necessaria all'interazione con il sistema operativo installato sui *device*; altri sono necessari in relazione al servizio offerto; altri, infine, risultano tecnicamente ridondanti**.

Le APP analizzate nello studio contengono, in totale, 266 permessi unici; la piattaforma per sviluppare APP in *Android* consente agli sviluppatori di prevedere anche nuove tipologie di permessi, così che i permessi sono caratterizzati da una certa dinamica per cui alcuni di essi vengono sostituiti nel corso del tempo. Occorre sottolineare che un numero considerevole di permessi riguarda aspetti tecnici legati al corretto funzionamento di una APP; se, ad esempio, si è interessati a un applicativo che misura percorsi stradali, è infatti necessario consentire l'accesso della APP ai rilevatori di posizione presenti nel *device*.

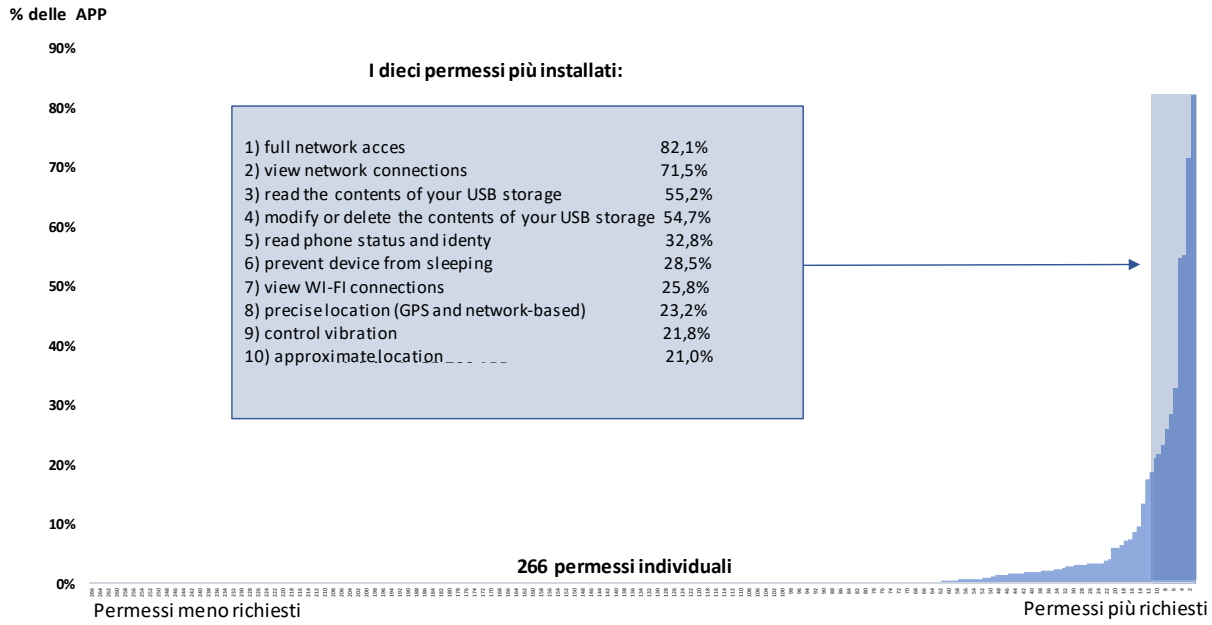


Figura 2.10 – Distribuzione dei permessi
Fonte: Elaborazioni AGCOM su dati *Google Play*

La **Figura 2.10** mostra la distribuzione dei permessi tra le APP: tra i 266 tipi di permessi solo 10 sono utilizzati da più del 20% delle APP, mentre un numero considerevole di permessi è utilizzato in pochissime applicazioni. Ben 20 permessi unici, ad esempio, vengono richiesti da una sola APP.

Ovviamente, l'interesse maggiore è per i permessi utilizzati dal maggior numero di applicazioni. Tra questi è importante individuare un criterio per **distinguere quelli che vengono richiesti con più probabilità al fine di acquisire dati individuali e quelli che, invece, sono tecnicamente necessari al buon funzionamento delle APP.** Al riguardo, in questo studio vengono utilizzate le più importanti categorizzazioni dei permessi presenti nella letteratura tecnico-economica.

In primo luogo, il *Pew Research Center*, in una ricerca condotta su un campione di APP disponibili su *Google Play Store*, distingue i permessi in due categorie: quelli riguardanti l'hardware e quelli concernenti i dati degli utenti (cd. *user info*).[106] Gli autori della ricerca evidenziano come la definizione di "*accesso alle informazioni degli utenti*" consenta di distinguere tra permessi che possono interferire con qualsiasi informazione dell'utente e quelli che non richiedono l'accesso a nessuna di esse.

Una seconda importante classificazione è quella elaborata dai ricercatori Kummer e Schulte (2016). I due studiosi, sulla base di una precedente classificazione effettuata da Sarma et al. (2012), individuano un numero limitato di permessi che possono risultare critici in termini di accesso a dati sensibili.[107] Essi costruiscono pertanto un indicatore che ricomprende una serie di permessi legati alla raccolta di dati degli utenti; tale variabile può a sua volta essere suddivisa in ulteriori quattro categorie, a seconda della tipologia delle informazioni acquisite: stato del *device*, localizzazione, attività di comunicazione e profilo.

L'individuazione di alcuni criteri per classificare i permessi più sensibili rispetto ai dati degli individui, tuttavia, di per sé non ha rilevanza; è necessario, infatti, considerare quante APP utilizzano questi specifici permessi e quanti *download* sono stati effettuati per comprenderne la diffusione tra gli utenti.

[106] Pew Research Center (2015), *Apps Permissions in the Google Play Store*, www.pewinternet.org
[107] KUMMER M. E., SCHULTE P., (2016); *When private information settles the bill: money and privacy in Google's market for smartphone applications*, ECONSTOR.
SARMA B.P., et al., (2012); *Android permissions: a perspective combining risk and benefits*, in *Proceedings of the 17th ACM symposium on Access Control Models and Technologies.*

Facendo riferimento ai 10 permessi più diffusi, la **Tabella 2.3** sintetizza, oltre al loro significato, anche se tali permessi possono essere considerati "sensibili" secondo le classificazioni del *Pew Research Center* e di Kummar & Schulte, ovvero se sono considerati come *dangerous* o *normal* sulla base della categorizzazione data direttamente dal gestore dello *store*, *Google*.

Tabella 2.3: Principali permessi per diffusione e rilevanza ai fini del trattamento dei dati sensibili

Permesso	Il permesso consente all'APP di…	Classificazione		
		Pew Center	Kummer & Schulte	*Google*
Accesso completo a Internet (*full network access*)	...creare prese di rete e utilizzare protocolli di rete personalizzati.	Sì	No	Dangerous
Vedere connessioni di rete (*view network connections*)	…vedere informazioni sulle connessioni di rete, ad esempio quali resti esistono e sono connesse	No	No	Normal
Accesso a memorie protette (*read the contents of your USB storage*)	..testare l'accesso a memorie USB o SD card che saranno disponibili su futuri device.	Sì	No	Normal
Modificare e cancellare i contenuti della propria memoria USB (*modify or delete the contents of your USB storage*)	…sovrascrivere la memoria USB o la SD card.	Sì	No	Dangerous
Leggere lo status e l'identità del telefono (*read phone status and identity*)	…accedere a caratteristiche telefoniche del device, quali il numero di telefono e l'ID del *device*.	Sì	Sì	Dangerous
Prevenire lo standby del device (*prevent device from sleeping*)	….prevenire lo standby/pausa del *device*.	No	No	Normal
Vedere connessioni Wi-Fi (*view WI-FI connections*)	….vedere informazioni sulle reti Wi-Fi (se sono attivate o quali sono i nomi dei device connessi).	Sì	No	Dangerous
Posizione esatta (*precise location GPS and network-based*)	…accedere alla posizione esatta usando il Global Positioning System (GPS) o le risorse dela rete come le torri cellulari e il Wi-Fi.	Sì	Sì	Dangerous
Controllo della vibrazione (*control vibration*)	…controllare la vibrazione del *device*.	No	No	Normal
Posizione approssimativa (*approximate location network-based*)	…accedere alla posizione approssimativa usando le risorse della rete come le torri cellulari e il Wi-Fi.	Sì	Sì	Dangerous
TOTALE PERMESSI "SENSIBILI"		**49**	**53**	**16**

Fonte: Elaborazioni AGCOM su dati *Google Play*

La **Tabella 2.3** si riferisce quindi ai permessi più diffusi tra le APP. Il lavoro di Kummer e Schulte individua come rilevanti per i dati digitali degli individui anche molti altri permessi, quali quelli che riguardano l'attività di comunicazione (es. consentire ad un'APP di leggere SMS o MMS, di registrare audio o di controllare le chiamate in uscita), e la profilazione dell'utente (consentire ad un APP di leggere il calendario, la rubrica e lo storico delle ricerche via *search web*).

2.7.2. Il valore dei dati individuali per imprese e consumatori

La successiva analisi ha riguardato il **prezzo delle applicazioni**, ossia il valore attribuito ad esse da imprese e consumatori, **dietro il quale si cela il valore attribuito ai dati transitati dal momento dell'acquisto della APP tramite il sistema dei permessi** descritto in precedenza.

Per quanto riguarda il **prezzo** delle APP (**Tabella 2.4**), ci si trova di fronte ad una distribuzione alquanto asimmetrica: l'86% delle applicazioni, infatti, può essere scaricata gratuitamente, mentre solo lo 0,5% (ossia 5.171 applicazioni) presenta un prezzo superiore ai 10 euro.

Tabella 2.4: Distribuzione delle APP per fascia di prezzo

Prezzo (€)	numero di APP	%
0	977.244	86,0
0-0,99	65.676	5,8
1-1,99	46.882	4,1
2-4,99	33.415	2,9
5-9,99	7.312	0,6
≥10	5.171	0,5
Totale	**1.135.700**	**100,0**

Fonte: Elaborazioni AGCOM su dati *Google Play*

Il fatto che la APP possa essere scaricata gratuitamente non preclude la possibilità che l'utente, in una fase successiva, decida di usufruire del servizio di *in-app purchase* che gli consente, a fronte di un pagamento, di ottenere servizi aggiuntivi (modalità *fremium*). Tuttavia, solo il 3% delle APP gratuite prevede un livello minimo, compreso tra 0,40 e 0,99 centesimi, di *in-app purchase*.

Il ***rating*** e il numero delle ***review*** rappresentano due variabili molto importanti nei processi decisionali degli utenti; al momento di effettuare lo scaricamento di un'APP, infatti, gli utenti sono in grado di aumentare il loro bagaglio informativo circa le specifiche tecniche e la qualità dell'APP attingendo utili informazioni proprio dai commenti e dal voto che gli utilizzatori del servizio hanno rilasciato. Gli APP *store* stessi consigliano di prendere visione di tali informazioni anche al fine di ridurre al minimo lo scaricamento di APP che potrebbero risultare dannose al corretto funzionamento del *device*.

Come osservato in precedenza, **la ricerca di un valore economico da poter attribuire allo scambio tra dati digitali e fruizione di un servizio (*download* di un'APP) passa attraverso lo studio dei permessi; a ciascun permesso, infatti, è possibile associare il rilascio di uno specifico insieme di dati legati all'individuo.**

Un primo interessante risultato emerge dalla **Tabella 2.5**; **le APP gratuite richiedono un numero significativamente maggiore di permessi rispetto a quelle a pagamento** (in media, circa il doppio, ossia 6,4 a fronte di 3,8). Il 50% delle APP a pagamento richiede fino a 3 permessi, per contro, il 50% di quelle gratuite ne richiede fino a 5.

Tabella 2.5: Numero medio di permessi

	n. di APP	% sul totale	n. medio di permessi
gratuite	977.244	86%	6,4
a pagamento	158.456	14%	3,8
Totale	**1.135.700**	**100%**	**6,0**

Fonte: Elaborazioni AGCOM su dati *Google Play*

Considerando i permessi più "sensibili", i risultati vengono confermati (**Tabella 2.6**): sia nel caso in cui si considerino le sole APP che richiedono almeno un permesso sensibile al trattamento dei dati individuali (**Panel B della Tabella 2.6**), sia nel caso più generale che comprende tutte le APP, cioè anche quelle che non richiedono dati sensibili, (**Panel A della Tabella 2.6**), il numero medio di permessi richiesti è decisamente maggiore quando le APP sono gratuite.

Tabella 2.6: Numero medio di permessi "sensibili"

Panel A: tutte le APP

	Pew Center	*Google*	Kummer & Schulte
gratuite	3,3	3,6	1,1
a pagamento	1,9	2,1	0,6

Panel B: APP che richiedono almeno un permesso "sensibile"

	Pew Center	*Google*	Kummer & Schulte
gratuite	8,1	7,1	9,7
a pagamento	6,1	5,4	7,2

Fonte: Elaborazioni AGCOM su dati *Google Play*

Una prima conclusione dello studio è quella secondo cui la gratuità di una APP presuppone il rilascio, attraverso il sistema dei permessi, di un maggior numero di dati digitali, in generale, e di quelli attinenti ai dati individuali, in particolare. Esiste in sostanza uno scambio implicito tra utenti e operatori web che incide sulla relazione commerciale primaria concernente la compravendita di APP.

Un ulteriore interessante aspetto riguarda la correlazione tra ***download*** e permessi. Il numero medio di *download* rappresenta, infatti, una valida approssimazione della domanda di APP da parte degli utenti, e, quindi, della potenziale quantità (**volume**) di dati raccolti dagli sviluppatori e dalle piattaforme di intermediazione (*APP store*). La **Figura 2.11** mostra la distribuzione delle APP per numero di *download* suddivise a seconda che si tratti di APP a pagamento o gratuite. Le due distribuzioni differiscono notevolmente, dal momento che, come d'altronde era facile supporre, le APP gratuite presentano un numero di *download* superiore. Più dell'80% delle APP a pagamento, invece, è scaricato da 1 a 100 volte, laddove, per quelle gratuite, si scende al 45%. Viceversa, quasi un terzo delle applicazioni gratuite è scaricato più di 1.000 volte, valore che è pari a meno del 4% per quelle a pagamento.

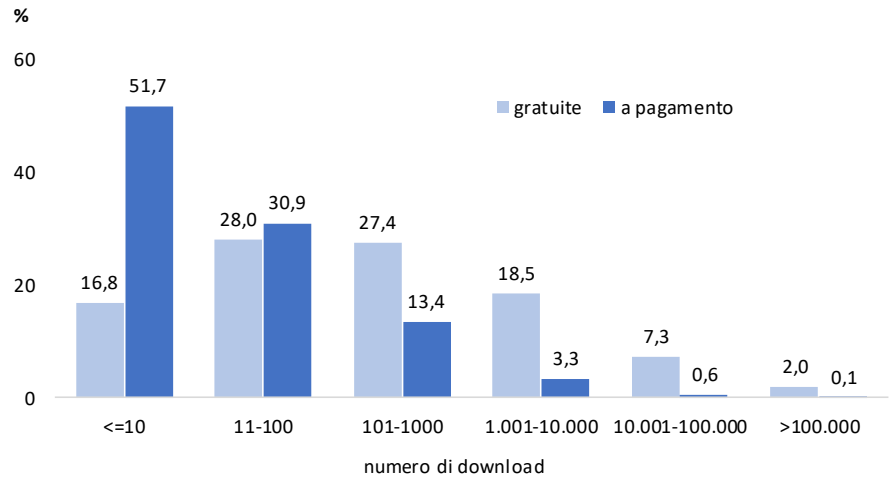


Figura 2.11 – Distribuzione delle APP per numero di *download*
Fonte: Elaborazioni AGCOM su dati *Google Play*

L'andamento dei *download* rileva un chiaro fenomeno di "coda lunga" anche nel mercato delle APP, come per il resto del web: in effetti, senza distinguere tra APP a pagamento e non, circa il 50% delle APP è scaricato meno di 100 volte e circa il 98% meno di 100.000 volte. **Questo determina che solo una manciata di APP, il 2%, risulta installata da un numero considerevole di utenti, in linea con quanto già mostrato in precedenza. Nel mondo, solo sei APP risultano installate più di 1 miliardo di volte: *Facebook*, *Google Gmail*, *Youtube*, *Google Maps*, *Google Search* e *Google Play Services*. Tale dato mostra ancora una volta come, a fronte di un numero elevatissimo di applicativi e operatori, il mercato sia in realtà concentrato in poche grandi piattaforme, che raggiungono numeri di diffusione non replicabili per gli altri operatori.**

Da questa analisi, seppure descrittiva, emergono **due tendenze rilevanti: *i)* il prezzo delle APP diminuisce al crescere del numero medio di permessi richiesti, anche se si considerano solo quelli sensibili ai dati individuali e *ii)* le APP scaricate più di frequente si caratterizzano per una maggiore presenza di permessi sensibili ai dati individuali.**

Un passo ulteriore nell'analisi è quello di verificare l'esistenza di una relazione causale al fine di verificare l'eventuale sussistenza di una relazione tra la domanda di APP degli utenti (*download*) e il numero di permessi richiesto dalle stesse applicazioni, nonché tra il prezzo delle APP fissato dagli sviluppatori e la quantità e la qualità di dati digitali transati.

In generale (v. Box 2), **si osserva una chiara significatività del numero di *download* ai permessi riguardanti informazioni sensibili**. In particolare, un aumento dei permessi considerati *user info* secondo la classificazione del *Pew Research*, comporta una riduzione del 5% dei *download*. Inserendo una classificazione più di dettaglio, come quella proposta da Kummer & Schulte, gli effetti dei permessi risultano più contrastanti; da un lato, infatti, quelli che richiedono l'accesso completo alla rete e al *device*, che si connotano per un taglio maggiormente tecnico, hanno un impatto positivo sui *download*; dall'altro lato, i permessi riguardanti le attività di comunicazione e la localizzazione dell'utente impattano invece in maniera negativa sul numero di *download*.

Per ciò che riguarda il prezzo, **i risultati confermano quanto già emerso dall'analisi delle statistiche descrittive; un modello di *business* che prevede la richiesta di un prezzo maggiore di zero, infatti, è associato ad una minor richiesta di permessi.** In tal senso, ad esclusione dei permessi che richiedono l'accesso allo stato del *device* dell'utente, l'effetto dei permessi, quando significativo, riduce sensibilmente la probabilità che un'APP sia a pagamento.

Dall'analisi empirica condotta su milioni di applicazioni, emerge un rilevante effetto del sistema dei permessi sottostanti al funzionamento di una APP, sia sulle scelte dei consumatori (*download*), sia sui modelli di *business* che le imprese intendono adottare. In particolare, **le relazioni che sono emerse evidenziano come il sistema dei permessi sia lo strumento attraverso il quale vengono scambiati dati tra imprese e consumatori.**

Tuttavia, **tale scambio non avviene nell'ambito di una transazione contrattuale certa in cui, tra l'altro, viene fissato il prezzo del prodotto, ma si sostanzia in uno scambio implicito, all'interno di una compravendita di altri servizi** (le APP). Ciò pone chiaramente enormi problemi circa l'efficienza del funzionamento dei mercati e di regolazione degli stessi.

BOX 2 – DOMANDA E OFFERTA DI APP: UN MODELLO ECONOMETRICO SULLO SCAMBIO DI DATI

Per quanto riguarda l'**analisi della domanda** di APP, si è proceduto alla stima di una regressione lineare che modella la domanda di APP, in termini di *download*, come funzione dei permessi che essa richiede, oltre che di una serie di variabili di controllo quali il prezzo, la categoria di appartenenza, il *rating* medio e il numero di *review* che l'applicazione ha.

La specificazione del modello stimato è la seguente:

$$Domanda_i = \alpha + \beta D_i + \theta X_i + \epsilon_i$$

dove la *Domanda* è rappresentata dal logaritmo del numero totale di *download* della generica APP *i*, mentre *X* include una serie di variabili di controllo (quali il prezzo, il totale dei permessi richiesti, il *rating* medio, il numero di *review* e categoria, lo sviluppatore dell'APP).

Il parametro β è quello di maggiore interesse poiché è associato a una variabile *dummy* (ossia binaria) che tiene conto del fatto che una APP richieda o meno il consenso ad un permesso sensibile ai dati individuali. Di conseguenza, se il parametro stimato risulta negativo, ciò implica che la presenza di permessi sensibili alle informazioni individuali determina una riduzione nella domanda.

Per la stima del modello è stato utilizzo un semplice OLS i cui risultati sono esposti nella tabella sottostante, mentre per la classificazione dei permessi si è utilizzata quella proposta nella TABELLA 2.1.

Stime OLS modello di domanda di APP

	Modello con classificazione permessi *Pew Research*	Modello con classificazione permessi *Google*	Modello con permessi *Hummer e Schulte*
Variabili di interesse			
Permessi *user information* (PEW)	-0.05***		
	(0.00)		
Permessi *dangeorus* (Google)		-0.01***	
		(0.00)	
Permessi che consentono l'accesso alla rete			0.07***
			(0.00)
Permessi che consentono di vedere la presenza di reti			-0.13***
			(0.00)
Permessi che consentono l'accesso allo stato del *device*			0.07***
			(0.00)
Permessi sulla localizzazione dell'utente			-0.05***
			(0.00)
Permessi attinenti all'attività di comunicazione			-0.10***
			(0.00)
Permessi riguardanti il profilo dell'utente			-0.02***
			(0.00)
Altri permessi			-0.06***
			(0.01)
Variabili di controllo	Sì	Sì	Sì
Categorie	Sì	Sì	Sì
F	126634.03	126441.59	111810.44
R-quadro	0.84	0.84	0.84
N	1,135,700	1,135,700	1,135,700

Errori robusti all'eteroschedasticità in parentesi. ***, **, * coefficienti significativamente differenti da 0 all'1%, al 5% e al 10%

Per quel che riguarda **l'analisi dell'offerta**, si è proceduto alla stima di un modello probabilistico in cui, tramite una variabile dicotomica uguale a 1 se una APP è a pagamento e uguale a 0 nel caso di un'applicazione gratuita, si analizza la scelta del modello di *business* da parte degli sviluppatori.

Il modello stimato è il seguente:

$$\Pr(Prezzo_i = 1) = \Lambda[\alpha + \beta D_i + \theta X_i + \epsilon_i]$$

Anche in questo caso, il principale parametro di interesse è rappresentato dal parametro β abbinato alla variabile *dummy* che individua se un permesso è da considerarsi o meno rilevante rispetto ai dati individuali: ci si attende che al crescere del numero di permessi richiesti dalla APP, la probabilità che l'applicazione presenti un prezzo maggiore di 0 si riduca. La tabella seguente riporta i risultati delle stime del modello probabilistico, in cui vengono riportate le precedenti classificazioni dei permessi.

Stime probabilistiche del modello di offerta

	Modello con classificazione permessi *Pew Research*	Modello con classificazione permessi *Google*	Modello con permessi *Hummer e Schulte*
Variabili di interesse			
Permessi *user information* (PEW)	-0.26***		
	(0.00)		
Permessi *dangeorus* (Google)		-0.67***	
		(0.00)	
Permessi che consentono l'accesso alla rete			-0.41***
			(0.01)
Permessi che consentono di vedere la presenza di reti			-0.55***
			(0.00)
Permessi che consentono l'accesso allo stato del *device*			0.09***
			(0.00)
Permessi sulla localizzazione dell'utente			-0.30***
			(0.01)
Permessi attinenti all'attività di comunicazione			0.01
			(0.01)
Permessi riguardanti il profilo dell'utente			-0.00
			(0.01)
Altri permessi			-0.03
			(0.02)
Variabili di controllo	Sì	Sì	Sì
Categorie	Sì	Sì	Sì
Pseudo R2	0.14	0.16	0.21
N	1,135,700	1,135,700	1,135,700

Errori robusti all'eteroscedasticità in parentesi. ***, **, * coefficienti significativamente differenti da 0 all'1%, al 5% e al 10%

2.7.3. L'inefficienza del sistema di scambio di dati

L'uso dei *big data* è diventato pervasivo estendendosi a un numero crescente di settori del sistema economico. I dati vengono raccolti e utilizzati per diverse finalità (cd. usi primari e secondari) e attraverso processi e tecnologie sempre più complessi e innovativi. In particolare, le piattaforme mobili e le relative componenti principali (*device*, sistema operativo, APP e APP *store*), configurandosi come strumenti personali di accesso a internet, svolgono un ruolo oramai primario sia nella vita quotidiana del cittadino sia, di conseguenza, nel processo di raccolta di dati online da parte degli operatori.

Il settore dei dati sconta, tuttavia, fenomeni di fallimento dei mercati. Come è già stato evidenziato in più occasioni dall'Autorità, infatti, i mercati in esame mostrano, a causa di fattori quali la presenza di forti esternalità di rete, una naturale tendenza alla concentrazione (con, al limite, situazioni monopolistiche del tipo *the winner takes all*).

In questo capitolo si è investigata **la relazione commerciale nello scambio, implicito ed esplicito, di servizi (APP) e dati tra utenti e fornitori di servizi internet**. In particolare, si è proceduto ad analizzare le caratteristiche delle *permission* che regolano la cessione di dati dagli utenti agli sviluppatori di APP nell'ambito di negozi virtuali (APP *store*). Attraverso un progetto di ricerca, sviluppato in collaborazione con il Dipartimento di Ingegneria Informatica Automatica e Gestionale dell'Università "La Sapienza" di Roma, il Servizio economico-statistico dell'Autorità ha proceduto ad analizzare i permessi e le caratteristiche di **milioni di APP contenute nel negozio virtuale di *Google*** (*Google Play*).

In un contesto in cui gli utenti manifestano, al contempo, un certo grado di consapevolezza nella cessione dei propri dati quando navigano in rete e un forte scetticismo sui rischi connessi alla gestione di tali informazioni proprio da parte degli operatori web, le relazioni che regolano questa cessione assumono una rilevanza centrale. Tuttavia, **lo scambio di dati individuali a fronte dell'offerta di servizi web spesso gratuiti, o comunque quasi sempre disponibili a un prezzo inferiore ai costi sottostanti, è implicito, ossia non è chiaramente e legalmente contrattualizzato, tanto che il mercato non assegna alcun prezzo alla transazione.**

L'uso dei dati per **fini non solo primari rende la struttura dell'ecosistema digitale profondamente diversa da quella degli altri media**, anch'essi parzialmente finanziati dalla raccolta pubblicitaria. In questo caso, infatti, **i dati vengono utilizzati non solo ai fini della vendita di contatti (personalizzati) agli inserzionisti pubblicitari, ma anche per un'ulteriore molteplicità di usi, anche sconosciuti al momento della raccolta**.

In questo contesto, **sono i dati stessi a rappresentare la principale merce di scambio degli agenti economici (utenti, operatori, *trader*, ecc.); tuttavia, il settore dei *big data* è strutturalmente incompleto tanto da mancare, almeno dal lato degli utenti, un meccanismo che ne disciplini la formazione del prezzo.**

L'analisi dell'Autorità ha approfondito la relazione commerciale implicita, che si sostanzia in pratica in una cessione da parte degli utenti di diritti sui propri dati a fronte, non già di un corrispettivo economico, ma dell'offerta di servizi web gratuiti, o comunque a prezzi contenuti, prossimi ai costi marginali.

Con una rigorosa analisi quantitativa, che ha riguardato milioni di APP, si è dimostrato che **tutti gli agenti economici - sia dal lato della domanda dei consumatori, sia da quello dell'offerta degli sviluppatori - scontano, nei propri comportamenti, l'assenza di un meccanismo istituzionale che regoli il commercio di dati.** Da un lato, infatti, i consumatori sono disposti a concedere i propri dati solo a fronte di minori prezzi delle applicazioni mobili; dall'altro lato, gli operatori web offrono le loro

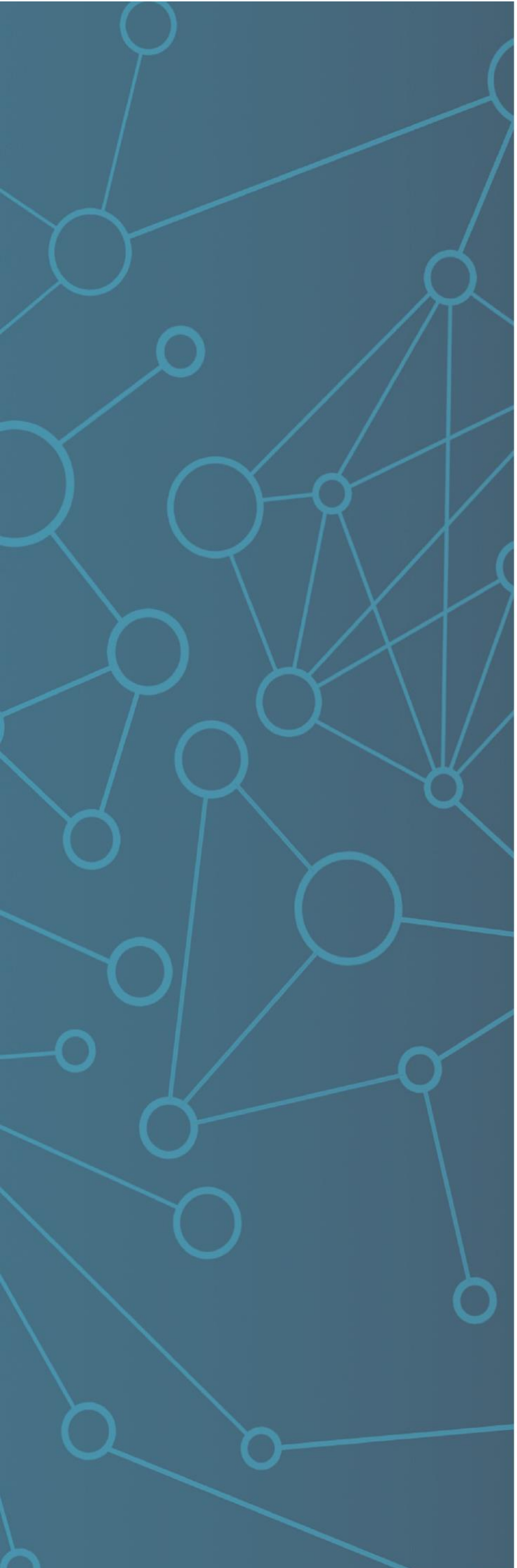
APP a prezzi minori, al limite nulli, solo a fronte dell'acquisizione di informazioni di dettaglio sugli utenti dei servizi.

In assenza di un mercato trasparente e di un meccanismo istituzionale che garantisca un quadro di regole stabili per gli attori coinvolti nella compravendita di dati, l'ecosistema digitale si è auto-regolato scontando l'incompletezza di questa transazione all'interno del prezzo dei servizi attraverso cui i dati vengono acquisiti dagli operatori e ceduti dagli utenti.

L'assenza di un vero e proprio meccanismo di mercato non può che rendere queste relazioni incomplete e inefficienti. In primo luogo, per quanto i permessi abbiano iniziato a distinguere le diverse tipologie di informazioni acquisite dall'operatore, attraverso una loro categorizzazione, **il consumatore non ha una chiara percezione di quali dati vengano ceduti e di come essi siano trattati, sia per gli usi primari, sia, a maggior ragione, per quelli secondari**. Si tratta, infatti, di **una transazione *una tantum* riguardante altri servizi (le APP) a fronte dell'uso dinamico e prolungato delle informazioni degli utenti.** È quindi la stessa configurazione strutturale del mercato e delle relative transazioni a essere distorta e, di conseguenza, a condurre a mercati incompleti che inevitabilmente falliscono.

Manca il principale strumento che regola, staticamente e dinamicamente, tutti i mercati (efficienti): il prezzo dei dati. Peraltro, essendo i dati fortemente eterogenei, il prezzo dovrebbe essere anche assai differenziato, a seconda della tipologia di informazioni scambiate.

L'assenza, inoltre, di corretti meccanismi di mercato tende a creare una **situazione di "sovrapproduzione" di dati:** l'allocazione delle risorse è socialmente inefficiente, non solo per l'assenza di sistemi che determinino contrattualmente tutte le specifiche dello scambio commerciale, ma, data la struttura dei prezzi impliciti, la quantità stessa del "bene" dati appare molto distante da quella ottimale dal punto di vista economico e sociale.



3

I BIG DATA NEL SISTEMA DELL'INFORMAZIONE

3.1. I big data, le piattaforme online e l'informazione

L'impiego di *big data*, come illustrato nei capitoli precedenti, è diventato pervasivo, estendendosi a un numero crescente di settori, dall'energia alla finanza, dalle assicurazioni all'*automotive*, dal commercio elettronico alla medicina, dall'intrattenimento all'informazione e così via. Grandi masse di dati, con differenti livelli di strutturazione, vengono raccolte e utilizzate per diverse finalità (usi primari e secondari), attraverso processi e tecnologie sempre più sofisticati; tale fenomeno riguarda, in particolare, gli operatori online, siano essi fornitori di servizi orizzontali (atti a soddisfare una pluralità di esigenze dell'utente) o verticali (atti a soddisfare bisogni specifici).

In questo contesto, l'individuo si configura quale fonte primaria di dati (cfr. Capitolo 3), conseguenza dell'interminabile scia di informazioni e tracce che ciascun utente lascia, più o meno consapevolmente, mentre svolge azioni online, per mezzo di dispositivi fissi e mobili. Si tratta di dati (immagini, video, parole digitate, *email*, …) riguardanti una moltitudine di caratteristiche e aspetti della vita dell'individuo, riconducibili alle proprie generalità, alla localizzazione, ai consumi, alle abitudini, agli interessi, alle ricerche effettuate, alle informazioni sociali (tra cui contatti, reazioni, contenuti condivisi, posizioni espresse).

Gli operatori che acquisiscono, direttamente e mediante *cookies* o altri sistemi di tracciamento, la quantità maggiore di dati, generati dagli individui durante la navigazione tramite *browser* o all'interno di applicazioni, sono senz'altro le piattaforme online. Le stesse, infatti, configurandosi come strumenti privilegiati di accesso a internet, svolgono ormai un ruolo primario sia nella vita quotidiana del cittadino sia, conseguentemente, nella raccolta di dati. Tra le piattaforme online e gli utenti si realizza, di fatto, uno scambio spesso implicito (v. Capitoli 3 e 4) per cui, a fronte delle informazioni rilasciate e generate dagli individui, le piattaforme online, grazie alla gestione e al processamento dei dati acquisiti, giungono ad offrire servizi gratuiti o a prezzi contenuti e a personalizzare l'esperienza dell'utente in rete, con l'obiettivo di massimizzare i propri profitti.

Più in generale, la capacità di accedere ai dati che riguardano i propri utenti e di utilizzarli come *asset* strategico secondo la logica dei mercati multi-versante costituisce un elemento pressoché costante nel modello di *business* delle piattaforme online.[108] Negli ultimi anni, utilizzi commerciali dei dati degli utenti stanno acquistando crescente importanza sia per le attività delle piattaforme sia per lo sviluppo dei servizi e dell'informazione online.

Motori di ricerca e *social network* sono, in questo senso, l'esempio più evidente di come molti servizi internet siano forniti senza un corrispettivo di prezzo per l'utente finale che, tuttavia, deve conferire, in maniera più o meno consapevole, alla piattaforma una serie di informazioni di carattere individuale per poter fruire il servizio.

Peraltro, l'utilizzo di *big data* da parte di motori di ricerca e *social network* rappresenta un aspetto di particolare importanza in ragione del ruolo sempre più rilevante svolto da queste piattaforme nel sistema dell'informazione, a livello internazionale e nazionale. Da un lato, le stesse, in virtù dei dati individuali di cui dispongono e che consentono un'accurata profilazione dell'utenza, si sono affermate come i *leader* mondiali nel settore della pubblicità online - risorsa che tuttora costituisce la fonte di finanziamento ampiamente prioritaria dell'informazione online -;[109] dall'altro, rappresentano ormai il veicolo distributivo

[108] Cfr. AGCOM (2014), *Indagine conoscitiva sui servizi internet e la pubblicità online*, https://www.agcom.it/indagine-conoscitiva-sul-settore-dei-servizi-internet-e-della-pubblicita-online.

[109] Cfr. AGCOM (2017), *Relazione annuale sull'attività svolta e sui programmi di lavoro*, pp. 120-128, https://www.agcom.it/documents/10179/3058729/RELAZIONE+ANNUALE+2017_documento+completo.pdf/2021e7ba-8250-4239-9a46-5d82fdbf702c; AGCOM (2015), *Indagine conoscitiva su Informazione e internet in Italia. Modelli di business, consumi*, professioni", https://www.agcom.it/documents/10179/1677802/Allegato+22-4-2015/69ae8f63-2301-46fd-a20b-f255546c5c42.

principale per l'informazione in rete, posto che la fruizione delle notizie su internet passa sempre più spesso attraverso motori di ricerca e *social network*.

Di conseguenza, **i *big data* assumono una valenza cruciale anche ai fini della tutela del pluralismo dell'informazione** dal momento che la disponibilità e il successivo utilizzo dei dati relativi ai consumatori, effettivi o potenziali, di servizi informativi sono diventati una leva competitiva essenziale per operare online, sia nel versante degli utenti sia in quello della raccolta pubblicitaria.[110]

L'acquisizione e l'impiego dei dati degli individui, infatti, si pongono alla base dei meccanismi stessi di funzionamento (come *crawling*, classificazione, associazione, prioritizzazione e filtraggio) delle piattaforme online che diffondono informazione. Al riguardo, motori di ricerca e *social network* vengono spesso definiti "fonti algoritmiche" di informazione appunto per richiamare la personalizzazione algoritmica, resa possibile dalla quantità e qualità dei dati raccolti sugli individui, che caratterizza i processi di generazione e divulgazione dei contenuti informativi sulle medesime piattaforme.

Più specificamente, la distribuzione dell'informazione su internet si sostanzia in un processo di disaggregazione e disintermediazione dell'offerta informativa tradizionale e di successiva riaggregazione e re-intermediazione da parte delle fonti algoritmiche. Pertanto, gli algoritmi sottostanti al loro funzionamento divengono decisivi nel determinare le modalità di fruizione dell'informazione da parte degli utenti, orientando significativamente il successo o meno, in termini di *audience*, di una notizia (o di un editore) rispetto a un'altra (sui meccanismi che presiedono il funzionamento degli algoritmi, cfr. anche il Box 5.1).

Sotto il profilo della domanda di informazione online, premesso che internet anche in Italia rappresenta ormai saldamente il secondo mezzo (dopo la televisione) più utilizzato per informarsi, le fonti algoritmiche sono costantemente presenti nella dieta informativa della maggior parte dei cittadini. I dati più recenti a disposizione dell'Autorità[111] (cfr. **Figura 3.1**) evidenziano come, nel 2017, il 54,5% degli italiani si informi attraverso strumenti governati da algoritmi, laddove il 39,4% della popolazione reperisce notizie direttamente da siti web e applicazioni degli editori (stampa quotidiana e periodica, radio e televisione, e testate esclusivamente online). Nel dettaglio, motori di ricerca e *social network* riscuotono preferenze analoghe tra la popolazione, raggiungendo entrambi una quota pari al 36,5% quando la finalità di fruizione è quella informativa.

Le fonti algoritmiche costituiscono, quindi, veri e propri *gatekeeper* per l'accesso all'informazione, "luoghi di passaggio" prediletti dagli utenti-cittadini, e, di riflesso, punto di approdo necessario per gli editori al fine di raggiungere i consumatori, incidendo chiaramente sulle strategie di distribuzione dei contenuti informativi attuate da questi ultimi, che, d'altra parte, rischiano, nel medio-lungo periodo, di perdere il contatto diretto con il pubblico e, quindi, la propria riconoscibilità a favore di quella dell'intermediario nell'ambito del quale avviene la fruizione dell'informazione online.

[110] Sul ruolo svolto dalle piattaforme online nel sistema informativo, l'Autorità sta conducendo anche la specifica Indagine conoscitiva *"Piattaforme digitali e sistema dell'informazione"*, volta ad analizzare la struttura e il funzionamento delle piattaforme nel diffondere informazione su internet, mettendo in luce le eventuali criticità sotto il profilo del pluralismo informativo. Inoltre, l'Autorità ha recentemente istituito il "*Tavolo tecnico per la garanzia del pluralismo e della correttezza dell'informazione sulle piattaforme digitali*", che ha l'obiettivo di promuovere l'autoregolamentazione delle piattaforme e lo scambio di buone prassi per l'individuazione e il contrasto dei fenomeni di disinformazione online (v. paragrafo 3.4).

[111] Cfr. AGCOM (2018), *Rapporto sul consumo di informazione*, cit.

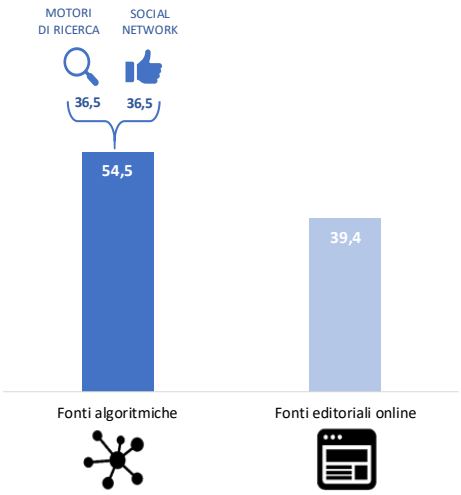


Figura 3.1 – Accesso all'informazione online da parte dei cittadini italiani (2017; % popolazione)
Fonte: elaborazioni AGCOM su dati GfK Italia

La prevalenza della frequenza di accesso alle fonti algoritmiche rispetto a quelle editoriali online per informarsi è ulteriormente rafforzata dall'importanza riconosciuta dagli individui alle prime, che, potendo ritenersi indice di una maggiore attenzione attribuita alla fruizione delle notizie divulgate, può essere considerata un'indicazione del consumo effettivo che gli utenti fanno delle predette fonti a scopo informativo.

In proposito, la **Figura 3.2** mostra come il 19,4% della popolazione indichi una fonte algoritmica come la più importante per informarsi. Spicca, in particolare, la rilevanza accordata a motori di ricerca (dal 9,7% dei cittadini) e *social network* (dal 6,8% degli italiani), che dopo i canali televisivi in chiaro nazionali e i quotidiani nazionali cartacei e digitali rappresentano rispettivamente la terza e la quarta fonte informativa più volte reputata come la più importante per informarsi, considerando la totalità dei mezzi di comunicazione classici e online.

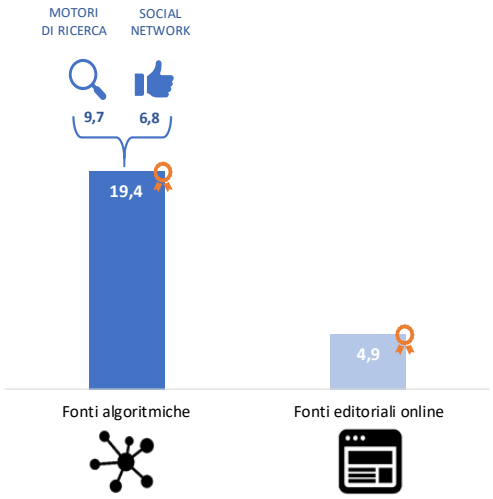


Figura 3.2 – Fonte di informazione ritenuta più importante dai cittadini italiani (2017; % popolazione)
Fonte: elaborazioni AGCOM su dati GfK Italia

In definitiva, **la diffusione dei *big data* sta alterando strutturalmente anche l'ecosistema informativo mondiale**. Come ha più volte evidenziato l'Autorità in questi anni, da un lato, l'avvento dei *data analytics* ha rivoluzionato il settore pubblicitario, che finanzia (in parte) le principali fonti informative

tradizionali e (in maniera preponderante) quelle online. In questo ambito, e in particolare nella pubblicità online, **le piattaforme, disponendo di un amplissimo spettro di dati e tecnologie di analisi, godono di un enorme e crescente vantaggio competitivo rispetto ad altri soggetti, quali gli editori, che sono attivi in rete**. Dall'altro, **le piattaforme di *big data* quali *social network* e motori di ricerca sono diventate la porta preferenziale dei cittadini per accedere all'informazione in rete**. Ciò non solo ha conseguenze per gli utenti, che consumano notizie e informazione, ma anche per gli editori, che con i loro prodotti informativi vogliono raggiungere gli utenti finali. I motori di ricerca e i *social network*, con i loro algoritmi (di indicizzazione, presentazione di contenuti, *newsfeed*, classificazione delle notizie, raccomandazione, ecc.; v. Box 3), rappresentano uno snodo fondamentale per chi produce e vuole distribuire informazione in rete. Il prossimo paragrafo è pertanto dedicato al ruolo svolto da queste piattaforme, e in particolare dai *social network*, nel sistema dell'informazione online.

BOX 3 – GLI ALGORITMI DELLE PIATTAFORME ONLINE E L'INFORMAZIONE

Gli algoritmi sono sottosistemi logici fondati su funzioni matematiche presenti in varie tipologie di software. Nel settore dell'informazione online, gli algoritmi si presentano come potenti strumenti utilizzati in particolare, ma non solo, per filtrare le notizie disponibili e presentarle agli utenti secondo un ordine, spesso personalizzato, derivante dall'applicazione di determinati criteri. In prima istanza, possiamo distinguere cinque categorie di algoritmi utilizzati dalle piattaforme online con riferimento all'ecosistema dell'informazione:

- *web search*: si tratta di algoritmi di indicizzazione e presentazione di contenuti aperti presenti sulle pagine web in formato HTML, con riferimento alle risposte alle *query* effettuate dagli utenti (tipici di *Google* – *PageRank* – e altri motori di ricerca);
- creazione di news *feed* sui *social network*: questo secondo tipo di algoritmi, tra cui rientra quello di *Facebook* (*EdgeRank*), determina la visibilità di ogni contenuto all'interno della piattaforma di *social networking* e la personalizzazione della pagina Notizie (la *news feed* appunto) di ogni utente, sulla base di determinate caratteristiche del post e dell'utente/pagina che lo pubblica;
- sistemi di raccomandazione: si tratta di algoritmi di selezione che suggeriscono determinati contenuti ai propri utenti in modo personalizzato;
- controllo e rimozione dei contenuti: sono algoritmi che intervengono in modalità automatica, effettuando una classificazione del post in base alla materia trattata e al suo orientamento;
- classificazione automatica di news: sono algoritmi tipicamente alla base del funzionamento di piattaforme che aggregano le notizie (es. *Google News*).

Gli algoritmi sono pertanto determinanti nel definire le modalità di consumo informativo degli utenti, e assumono un valore significativo anche dal lato dell'offerta nell'orientare il successo di talune notizie (ed editori) rispetto ad altre e nel determinare le scelte di editori e giornalisti. Di conseguenza, cambiamenti dei principali algoritmi possono modificare profondamente l'ecosistema dell'informazione online, sia dal lato della domanda che da quello dell'offerta.

3.2. Il ruolo dei social network nel sistema dell'informazione

Tra tutte le piattaforme online, **i *social network*** – in ragione del tempo trascorso dagli utenti all'interno degli stessi, delle molteplici azioni che gli individui compiono e reazioni che esprimono attraverso i propri profili/pagine/*account*, nonché delle relazioni sociali che instaurano – **si configurano certamente tra gli operatori in grado di acquisire la maggiore varietà e il maggior volume di dati (ossia di *big data*, in tal senso v. Capitolo 1) sugli individui, compresi quelli relativi alle preferenze ideologiche e politiche e ai contenuti informativi letti, visualizzati, graditi, commentati e condivisi.**

In aggiunta, all'interno del sistema informativo, i *social network* si contraddistinguono per collocazione e caratteristiche peculiari rispetto a tutte le altre fonti di informazione per almeno tre motivi principali.

1) Come argomentato nel paragrafo precedente, il funzionamento dei *social network*, anche per quel che concerne la proposizione di contenuti a carattere informativo, è governato da algoritmi in grado di filtrare costantemente, secondo criteri predeterminati, le notizie disponibili e presentarle agli utenti secondo un ordine generalmente personalizzato, ossia che tiene conto della tipologia di utente.

2) In uno scenario caratterizzato dallo "spacchettamento" del prodotto informativo e da una fruizione frammentata dei contenuti (articoli, commenti, video, *post*, ecc.), i *social network* fungono da intermediari per l'accesso dei cittadini all'informazione, accesso che, molto spesso, è frutto anche dell'incidentalità e casualità della scoperta delle notizie da parte dell'utente.

3) I *social network*, infine, consentono l'ingresso nell'ecosistema informativo di fonti estranee al classico circuito dell'informazione, attraverso profili di utenti comuni, pagine/*account* di informazione non professionali, pagine/*account* satirici, ecc. Nella maggior parte dei casi, peraltro, sui *social network*, i contenuti informativi a carattere giornalistico e quelli generati dagli utenti e da altre figure non professionali assumono la stessa rilevanza, dal momento che la selezione e prioritizzazione delle informazioni da mostrare agli utenti avviene con meccanismi di aggiornamento automatici, sulla base di dati come la prossimità dei contenuti, i *post* degli amici, le reazioni, le condivisioni e i commenti, piuttosto che sulla credibilità e qualità giornalistica o il rilievo in termini di pubblico interesse del contenuto/notizia.

I *social network*, dunque, rappresentano uno spazio virtuale in cui l'utente è posto in comunicazione diretta con tutti gli attori presenti, con varie funzioni, nel sistema informativo: dai giornalisti agli editori, dai politici agli *influencer*, dalle figure non professionali che diffondono notizie agli altri utenti (amici e amici degli amici).

Sui *social network*, ciascun individuo da mero fruitore di notizie diviene, almeno in potenza, parte attiva nella diffusione di informazioni, opinioni e punti di vista, potendo partecipare a tale processo con diversi gradi di intensità, a seconda delle azioni che sceglie di compiere. In tal senso, l'utente può non limitarsi a cliccare sul *link* di una notizia, ma esprimere una reazione rispetto alla stessa, condividerla, commentarla, fino a partecipare a una discussione sulla notizia e postare proprie immagini, foto e video in merito all'argomento. Tutte azioni, queste ultime, di per sé atte a favorire la diffusione delle notizie e innescare fenomeni di "viralizzazione" dei contenuti informativi.

A livello internazionale, **i *social network* sono definitivamente divenuti parte integrante della dieta informativa quotidiana dei cittadini. In gran parte del mondo, infatti, sono considerati tra le fonti utilizzate più assiduamente per reperire ogni genere di notizia, rivestendo una grande e crescente valenza nel processo di formazione dell'opinione pubblica, e quindi sotto il profilo del pluralismo informativo** (al riguardo, cfr. anche il paragrafo successivo).

Un recente studio del *Pew Research Center*, condotto su 38 paesi, evidenzia come, nel 2017, l'accesso giornaliero (ossia, effettuato una o più volte al giorno) ai *social network* allo scopo di informarsi coinvolga

mediamente oltre un terzo della popolazione adulta (cfr. **Figura 3.3**), e tale valore raggiunge il 48% se si considerano anche coloro che dichiarano di informarsi attraverso i *social network* meno spesso di tutti i giorni[112]. La **Figura 3.3** mostra, peraltro, come nel complesso non si registrino differenze significative tra i Paesi economicamente avanzati (in cui il ricorso quotidiano ai *social network* per acquisire informazioni è mediamente pari al 36%) e quelli in via di sviluppo (in cui in media la popolazione adulta si informa giornalmente attraverso i *social network* nel 33% dei casi). Focalizzando l'attenzione sulle singole realtà nazionali, in 3 dei 38 Paesi analizzati (Corea del Sud, Libano e Argentina) a informarsi tutti i giorni sui *social network* è più della metà dei maggiorenni e altri 7 paesi (Canada, Australia, Svezia, Vietnam, Turchia, Cile e Brasile) esibiscono percentuali di utilizzo superiori al 40%. Soltanto in 4 Paesi (Indonesia, Senegal, India e Tanzania), invece, il tasso di fruizione quotidiana dei *social network* a scopo informativo è inferiore al 20%.

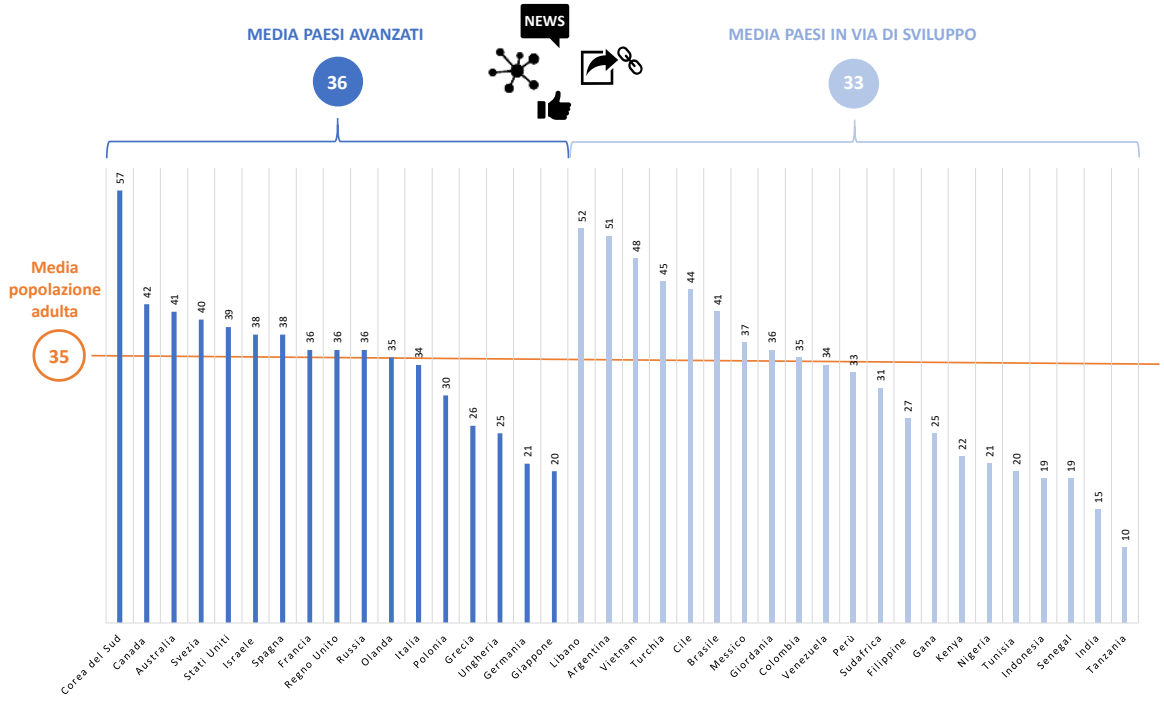


Figura 3.3 – Utilizzo dei *social network* per informarsi quotidianamente (2017; % popolazione di 18 anni e più)

Fonte: *Pew Research Center*, *Global Attitudes Survey 2017*

Come si è avuto modo di anticipare (cfr. **Figura 3.2**), il ruolo dei *social network* nel soddisfare la domanda di informazione da parte della popolazione italiana è ulteriormente avvalorato dalla primaria importanza dichiaratamente riconosciuta loro dagli individui, nonostante quella informativa sia soltanto una delle molteplici finalità che possono guidarne la fruizione. Il recente studio dell'Autorità sul consumo di informazione in Italia[113], che largo spazio ha dedicato anche all'accesso e al consumo di informazione politica, rileva altresì come il 14,9% degli elettori italiani affidi ai *social network* la propria ricerca di informazione ai fini delle scelte politico-elettorali (cfr. **Figura 3.4**). Tale valore, che posiziona i *social network*, primi tra le fonti informative online, al terzo posto (soltanto dopo canali televisivi e quotidiani nazionali) nella graduatoria complessiva dei mezzi utilizzati per informarsi di politica, corrisponde al 43,8% del totale degli elettori che dichiarano di reperire online notizie sulla politica.

[112] Cfr. Pew Research Center (2018), *Publics Globally Want Unbiased News Coverage, but Are Divided on Whether Their News Media Deliver*. Lo studio si basa sui risultati della *Global Attitudes Survey 2017*.
[113] Cfr. AGCOM (2018), *Rapporto sul consumo di informazione*, cit.

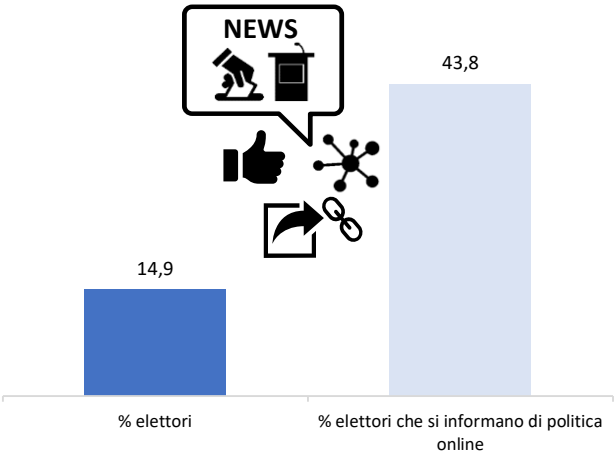


Figura 3.4 – Utilizzo dei *social network* per informarsi sulle scelte politico-elettorali in Italia (2017; %)
Fonte: elaborazioni AGCOM su dati GfK Italia

Nonostante il rilievo attribuito dai cittadini ai *social network* quali strumenti di informazione, comunicazione e relazione sociale, gli stessi, negli ultimi anni, sono stati posti al centro del dibattito internazionale sulla diffusione di forme patologiche quali quelle relative alla polarizzazione[114], che innesca il formarsi di bolle ideologiche (o *echo chamber*)[115] in rete e, più in generale, di fenomeni di disinformazione.

In proposito, diversi studi scientifici fondati sull'analisi di milioni di dati derivanti dalla fruizione dei *social network* esaminano il ruolo esercitato dalla polarizzazione ideologica e dai meccanismi di *confirmation bias* (ossia, la tendenza ad acquisire informazioni coerenti alle proprie preferenze ideologiche a scapito di quelle contrastanti) nella disseminazione della disinformazione online. Tra questi si annoverano, in particolare, i lavori che provano la centralità assunta da forme di esposizione selettiva ai contenuti informativi sui *social network* nella creazione e diffusione di disturbi dell'informazione[116]. I ricercatori giungono, infatti, a stabilire l'esistenza di un legame diretto tra il livello di polarizzazione degli argomenti trattati e la tendenza di questi ultimi a divenire oggetto di disinformazione[117]. Gli stessi riscontrano, peraltro, come la tendenza alla polarizzazione da parte degli utenti attorno a tematiche dibattute (quali,

[114] Nell'ambito del già richiamato Rapporto dell'Autorità sul consumo di informazione, l'analisi della relazione sussistente tra la polarizzazione ideologica degli utenti dei *social network* e le loro attività svolte in rete ha mostrato come la polarizzazione possa avere un effetto significativo sul maggior impegno (*engagement*) nei confronti delle notizie divulgate dai *social network*. Il legame tra lo svolgimento di tutte le azioni informative sui *social network* (incluse quelle a più alto tasso di coinvolgimento dell'utente) e la polarizzazione ha evidenti riflessi sul concretizzarsi di fenomeni di diffusione di posizioni radicalizzate e creazione di bolle ideologiche.

[115] Le *echo chamber* sono caratterizzate da individui che discutono solo all'interno di una cerchia di persone vicine ideologicamente. Le stesse, pertanto, tendono a ricalcare e acuire le problematiche di esposizione selettiva e *confirmation bias*.

[116] Cfr., tra gli altri, BESSI A., COLETTO M., DAVIDESCU G.A., SCALA A., CALDARELLI G., QUATTROCIOCCHI W. (2015), "Science vs Conspiracy: Collective Narratives in the Age of Misinformation", *PLoS ONE* 10(2); ZOLLO F., KRALJ NOVAK P., DEL VICARIO M., BESSI A., MOZETIČ I., SCALA A., CALDARELLI G., QUATTROCIOCCHI W., (2015), "Emotional Dynamics in the Age of Misinformation", *PLoS ONE* 10(9); BESSI A., PETRONI F., DEL VICARIO M., ZOLLO F., ANAGNOSTOPULOS A., SCALA A., CALDARELLI G., QUATTROCIOCCHI W. (2016), "Homophily and Polarization in the Age of Misinformation", *The European Physical Journal Special Topics*, 225(10); DEL VICARIO M., BESSI A., ZOLLO F., PETRONI F., SCALA A., CALDARELLI G., STANLEY H.E., QUATTROCIOCCHI W., (2016), "The Spreading of Misinformation Online", *Proceedings of the National Academy of Science* 113(3); ZOLLO F., BESSI A., DEL VICARIO M., SCALA A., CALDARELLI G., SHEKHTMAN L., HAVLIN S., QUATTROCIOCCHI W., (2017), "Debunking in a World of Tribes", *PLoS ONE* 12(7); e SCHMIDT A.L., ZOLLO F., CALDARELLI G., SCALA A., QUATTROCIOCCHI W., et al., (2017), "Anatomy of News Consumption on *Facebook*", *Proceedings of the National Academy of Sciences*, 114(12), pp. 3035-3039.

[117] M. DEL VICARIO, W. QUATTROCIOCCHI, A. SCALA, F. ZOLLO (2018), "Polarization and Fake News: Early Warning of Potential Misinformation Targets", *arXiv preprint arXiv:1802.01400*.

ad esempio, la Brexit[118] o il referendum costituzionale italiano del 2016[119]) si manifesti con modalità analoghe su *social network* diversi (che raggiungono *target* di utenti dissimili), basati su algoritmi differenti.

Per mezzo dei *social network*, dunque, i sistemi di personalizzazione automatica (che operano sulla base dei *big data* acquisiti), da un lato, e le azioni di condivisione di contenuti informativi compiute dagli utenti, dall'altro, facilitano la proliferazione di notizie false e la propagazione virale dei contenuti.

Al riguardo, Vosoughi, Roy e Sinan (2018)[120] hanno studiato le modalità e la velocità di diffusione di notizie false attraverso i *social network*, confrontandole con le caratteristiche di propagazione di notizie vere[121]. Nello specifico, la ricerca esamina notizie vere e false diffuse su Twitter dal 2006 al 2017 e classificate come "vere" o "false" sulla base delle informazioni di 6 organizzazioni indipendenti di *fact-checking*. Le risultanze prodotte dall'analisi di 126.000 storie *twittate* da 3 milioni di persone più di 4,5 milioni di volte rivelano come, per tutte le categorie di notizie considerate (politica, economia, guerre e terrorismo, disastri ambientali, scienza e tecnologia, intrattenimento, curiosità), quelle false siano in grado di diffondersi più velocemente, in maniera più estesa e con maggiore pervasività rispetto alle notizie vere (cfr. **Figura 3.5**).

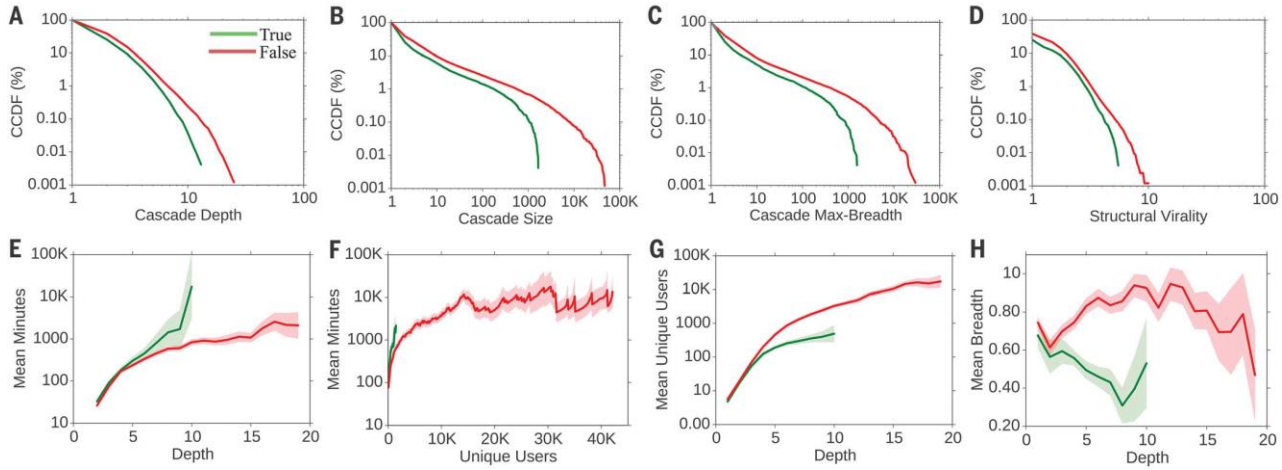


Figura 3.5 – Modalità di diffusione di notizie vere e false su Twitter

Fonte: *Science*, 2018, Vol. 359, pp. 1146-1151

Peraltro, tali effetti, come rappresentato nella **Figura 3.6**, diventano più pronunciati quando si tratta di notizie false inerenti alla politica, laddove notizie false più popolari mostrano dinamiche di diffusione più ampie (raggiungono il maggior numero di persone) e accelerate (in termini di velocità di propagazione e viralità).

[118] Cfr. DEL VICARIO M., F. ZOLLO, CALDARELLI G., SCALA A., QUATTROCIOCCHI W., (2017), "Mapping Social Dynamics on *Facebook*: The Brexit Debate", *Social Networks*, Vol. 50, pp. 6-16.

[119] DEL VICARIO M., GAITO S., QUATTROCIOCCHI W., ZIGNANI M., ZOLLO F., (2017), "News Consumption during the Italian Referendum: A Cross-Platform Analysis on *Facebook* an Twitter", *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, http://ieeexplore.ieee.org/document/8259827.

[120] Cfr. VOSOUGHI S., ROY D., ARAL S., (2018), "The spread of true and false news online", *Science*, Vol. 359, pp. 1146-1151.

[121] Altri studi, in precedenza, si erano soffermati sulle modalità di diffusione di determinate tipologie di notizie false sui *social network*. Cfr., tra gli altri, BESSI A., COLETTO M., DAVIDESCU G.A., SCALA A., CALDARELLI G., QUATTROCIOCCHI W. (2015), "Science vs Conspiracy: Collective Narratives in the Age of Misinformation", cit.; e DEL VICARIO M., BESSI A., ZOLLO F., PETRONI F., SCALA A., CALDARELLI G., STANLEY H.E., QUATTROCIOCCHI W., (2016), "The Spreading of Misinformation Online", cit.

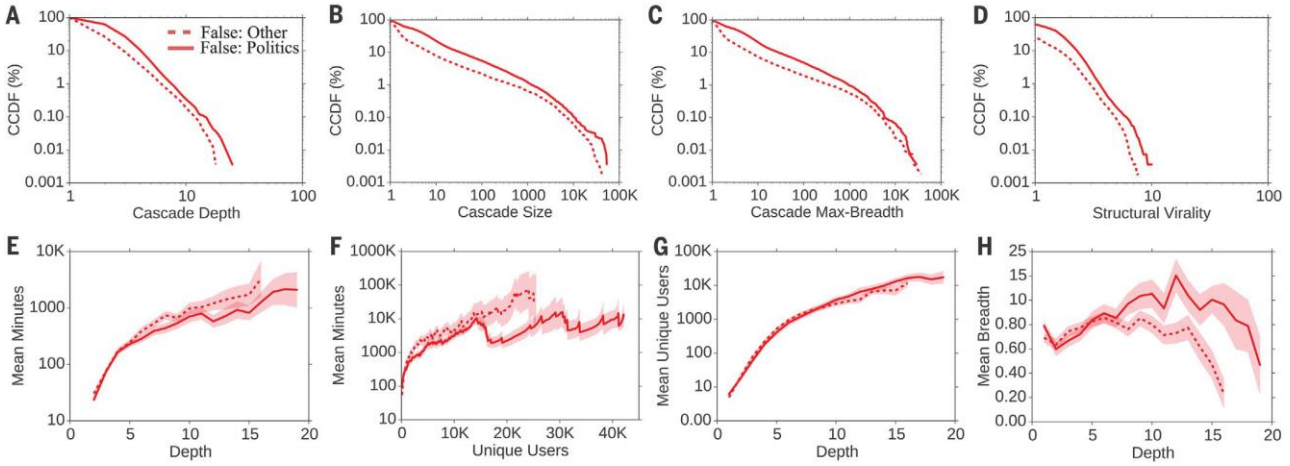


Figura 3.6 – Modalità di diffusione di notizie false di politica rispetto alle altre su Twitter

Fonte: *Science*, 2018, Vol. 359, pp. 1146-1151

3.3. L'influenza dei social network sulla formazione dell'opinione pubblica

In considerazione della crescente rilevanza dei *social network* per informarsi e, al contempo, degli aspetti di criticità sotto il profilo della tutela del pluralismo informativo che – secondo le dinamiche sopra esposte – ne possono accompagnare la fruizione, diverse branche della letteratura si sono interrogate sull'effettiva capacità dei *social network*, reti sociali virtuali, di influire sui comportamenti reali (*offline*). La tematica assume particolare rilievo all'interno di un'analisi sui *big data*, vista la crucialità che l'acquisizione e l'utilizzo dei dati degli individui rivestono per il funzionamento stesso dei *social network* anche con riferimento alla distribuzione dei contenuti informativi.

La produzione di lavori scientifici sull'argomento, sia teorici che empirici, risulta già copiosa e variegata seppur in continua espansione, alla luce della repentina evoluzione tecnologica e delle informazioni analizzabili generate dalle tracce digitali lasciate dagli utenti durante la navigazione sui *social network*[122]. Ai fini dello studio condotto nel presente Rapporto, si ritiene opportuno, quindi, passare in rassegna le evidenze salienti emerse dalla ricerca più recente circa l'influenza esercitata dai *social network* sul processo con cui gli individui costruiscono la propria visione della realtà. In ragione delle competenze dell'Autorità in materia, l'accento viene posto in particolare **sugli effetti prodotti dai *social network* sulla formazione dell'opinione pubblica e le scelte politiche degli utenti, tenendo presente che l'attitudine dei *social network* a raggiungere grandi masse di individui comporta che anche piccoli effetti possano generare cambiamenti comportamentali per migliaia di persone, così da poter incidere sugli esiti elettorali. Un successivo Rapporto dell'Autorità analizzerà, grazie alla collaborazione di eminenti ricercatori in materia (e in particolare del Prof. Walter Quattrociocchi[123]), tutti i processi antecedenti l'effettiva scelta elettorale del cittadino, ossia quelli relativi alla formazione dell'opinione pubblica in rete**.

Un aspetto preliminare da considerare, in quanto di per sé idoneo a innescare forme di "contagio" che potenzialmente possono riguardare tutta la sfera delle percezioni umane, incluse le sensazioni positive e

[122] A tale riguardo, vale segnalare, tra l'altro, l'affermarsi di un nuovo campo di ricerca, la scienza sociale computazionale, che studia i fenomeni sociali associati al consumo dei *social network*, adottando un approccio quantitativo (fondato sull'utilizzo di grandi masse di dati) e multidisciplinare (che coinvolge ambiti quali la matematica, la statistica, la fisica, la sociologia, l'informatica); v. LAZER D., (2009), "Computational Social Science", *Science*, Vol. 323, pp. 721-723.

[123] V. esiti della procedura comparativa per l'affidamento di un incarico di ricerca su "Informazione e piattaforme digitali", cfr. https://www.agcom.it/documents/10179/8623861/Allegato+25-1-2018/e2200786-4963-4d19-a7e0-a437d24061c5?version=1.0.

negative nei confronti di un argomento/esponente politico, è la capacità dei *social network* di influire sugli stati emotivi. Già nel 2014, Kramer, Guillory e Hancock[124] rilevano che gli stati emotivi, sui *social network*, possono essere trasferiti agli altri attraverso il contagio emozionale, portando persone diverse a provare le medesime emozioni senza la propria consapevolezza. Nel dettaglio, gli autori conducono un esperimento su 689.003 utenti di *Facebook*, modificando la quantità di contenuto emotivo mostrato nel *news feed* degli individui. L'esperimento evidenzia che in presenza di una riduzione delle espressioni positive, le persone producono meno *post* positivi e più *post* negativi; viceversa, in presenza di una riduzione di espressioni negative, si verifica lo schema opposto. Questi risultati suggeriscono che le emozioni espresse da altri sul *social network* sono suscettibili di influire sulle emozioni degli utenti, costituendo prove sperimentali di un contagio su vasta scala, pur in assenza di interazione personale (è sufficiente l'esposizione allo stato di un amico che esprime un'emozione), rimarcando anche da questo punto di vista l'estrema rilevanza della selezione di contenuti mostrati sulle piattaforme.

Analogamente, Coviello et al. (2014)[125], a partire dai dati di milioni di utenti di *Facebook*, mostrano che il contenuto emotivo dei messaggi di stato degli utenti è in grado di influire sui messaggi di stato degli amici, riscontrando come i *social network* possano amplificare l'intensità della sincronia emotiva globale.

Il trasferimento di emozioni, sensazioni e, più in generale, percezioni (anche ideologiche) sui *social network* sono stati oggetto di studio anche da parte di Jost et al. (2018)[126]. Gli autori, passando in rassegna una varietà di lavori sui movimenti di protesta negli Stati Uniti, in Spagna, in Turchia e in Ucraina, rilevano come i *social network* possano concretamente esercitare un ruolo importante per lo scambio di informazioni e il coordinamento dell'azione collettiva. E ciò non solo in quanto le informazioni rilevanti per il coordinamento delle attività di protesta (come le notizie su trasporti, affluenza, presenza della polizia, pericoli, servizi medici, assistenza legale) sono atte a diffondersi rapidamente ed efficientemente attraverso i *social network*, ma anche perché le piattaforme sociali agevolano la trasmissione di messaggi emotivi e motivazionali, sia a sostegno che in opposizione all'attività di protesta, in grado di enfatizzare determinati stati emotivi (quali l'indignazione morale, l'identificazione sociale, l'appartenenza al gruppo e le preoccupazioni su equità, giustizia sociale, privazione), nonché temi esplicitamente ideologici, che, diffondendosi su larga scala, possono incidere fino a determinare il successo o il fallimento dei movimenti di protesta stessi.

Inoltre, Bond e Messing (2015)[127], analizzando i dati di oltre 6 milioni di utenti di *Facebook*, indagano la relazione tra l'esposizione al disaccordo sul *social network* da parte di un individuo e l'effettivo esercizio del voto. In sostanza, dalla ricerca emerge che un aumento della distanza ideologica rispetto ai propri amici è associato a tassi più bassi di affluenza alle urne da parte dell'utente considerato.

L'influenza sociale scaturente dalla fruizione dei *social network* sul comportamento degli elettori è stata, altresì, oggetto di un vastissimo studio sperimentale, condotto in occasione delle elezioni del Congresso tenutesi il 2 novembre 2010 negli Stati Uniti[128]. Bond et al. (2012), per verificare l'ipotesi che il comportamento politico possa diffondersi attraverso un *social network*, hanno compiuto un esperimento

[124] KRAMER A.D.I., GUILLORY J.E., HANCOCK J.T., (2014), "Experimental evidence of massive-scale emotional contagion through social networks", *PNAS*, Vol. 111, N. 24, pp. 8788-8790.
[125] COVIELLO L., SOHN Y., KRAMER A.D.I., MARLOW C., FRANCESCHETTI M., CHRISTAKIS N.A., FOWLER J.H. (2014), "Detecting Emotional Contagion in Massive Social Networks", PLoS ONE 9(3).
[126] JOST J.T., BARBERP, BONNEAU R., LANGER M., METZGER M., NAGLER J., STERLING J., TUCKER J.A., (2018), "How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks Advances", *Political Psychology*, Vol. 39, Suppl. 1.
[127] BOND R.M., MESSING S., (2015), "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on *Facebook*", *American Political Science Review*, Vol. 109, N. 1, pp. 62-78.
[128] BOND R. M., FARISS C.J., JONES J. J., KRAMER A.D.I., MARLOW C., SETTLE J. E., FLOWER J. H., (2012), "A 61-million-person experiment in social influence and political mobilization", *Nature*, Vol. 489, pp. 295-298.

sugli utenti di almeno 18 anni che hanno avuto accesso a *Facebook* il giorno delle elezioni. Gli utenti sono stati assegnati in modo casuale a 3 differenti gruppi:

- un gruppo, al quale è stato mostrato un **"messaggio social"**, composto da oltre 60 milioni di persone che, il 2 novembre 2010, hanno visualizzato nella parte superiore del *news feed* di *Facebook* un messaggio che incoraggiava l'utente a votare, conteneva un *link* per la ricerca dei seggi locali, un pulsante cliccabile con la scritta "Ho Votato", un contatore che indicava quanti altri utenti di *Facebook* avevano già votato, oltre a riportare fino a 6 immagini del profilo di amici di *Facebook* dell'utente selezionati in modo casuale tra coloro che avevano già cliccato sul pulsante "Ho Votato" (cfr. **Figura 3.7**);

Figura 3.7 – Messaggio social mostrato durante l'esperimento di Bond et al.

Fonte: *Nature*, 2012, Vol. 489, pp. 295-298

- un gruppo, al quale è stato mostrato un **"messaggio informativo"**, composto da 611.044 utenti, che hanno visualizzato un messaggio che incoraggiava a votare, riportava il *link* per la ricerca dei seggi locali, conteneva il pulsante cliccabile con la scritta "Ho Votato" e il contatore che indicava quanti altri utenti di *Facebook* avevano già votato. A differenza del "messaggio social", il "messaggio informativo" non mostrava alcun volto di amici che avevano votato (cfr. **Figura 3.8**);

Figura 3.8 – Messaggio informativo mostrato durante l'esperimento di Bond et al.

Fonte: *Nature*, 2012, Vol. 489, pp. 295-298

- un gruppo di controllo, composto da 613.096 utenti, ai quali non è stato mostrato alcun messaggio nella parte superiore del proprio *news feed*.

All'esito dell'esperimento, gli autori hanno esaminato sia gli effetti diretti che la proposizione dei messaggi ha generato sul comportamento degli utenti di ciascun gruppo, sia gli effetti indiretti prodotti nei riguardi di amici e amici degli amici degli utenti sottoposti all'esperimento.

Per quel che concerne gli effetti diretti (cfr. **Figura 3.9**), Bond et al. hanno riscontrato negli utenti esposti al "messaggio social" una maggiore propensione all'autodichiarazione di voto, alla ricerca di notizie sui seggi locali, nonché all'effettiva partecipazione al voto (attestata attraverso l'esame dei registri pubblici).

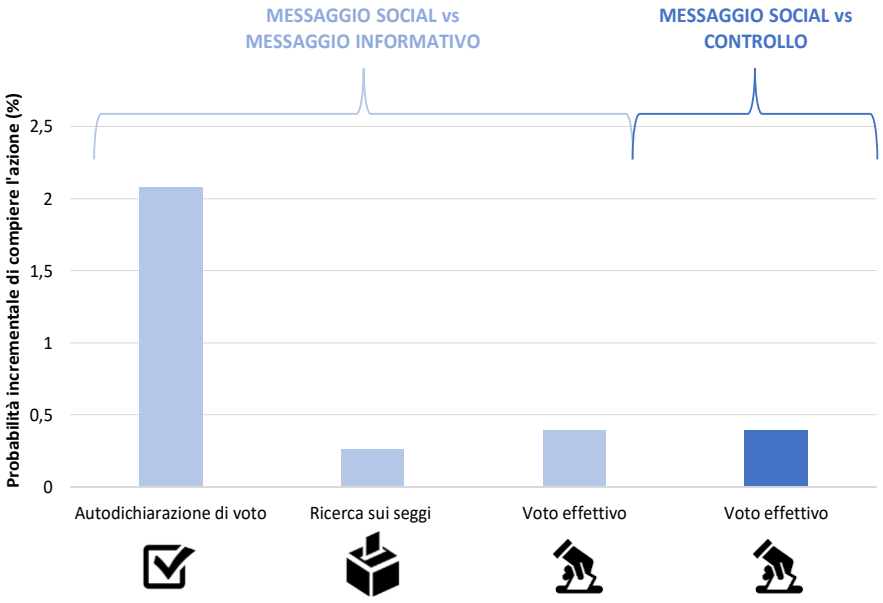


Figura 3.9 – Effetti diretti dell'esposizione ai messaggi sulle azioni politiche dell'utente

Fonte: *Nature*, 2012, Vol. 489, pp. 295-298

Per studiare gli effetti indiretti, gli autori hanno ricostruito la composizione delle reti amicali degli utenti del campione (costituita in media da 149 amici), distinguendo le relazioni caratterizzate da legami forti da quelle caratterizzate da legami deboli, in base ai livelli di interazione valutati per ciascuna coppia di soggetti. L'effetto del trattamento sperimentale per ogni amico è stato, poi, misurato confrontando il comportamento degli amici connessi a un utente esposto al "messaggio social" con il comportamento degli amici connessi a un utente del gruppo di controllo. I risultati ottenuti hanno rilevato come gli effetti del trattamento osservato su un amico aumentino all'aumentare della forza del legame con l'utente esposto al messaggio. In particolare, i legami forti (sussistenti tra gli amici più intimi) sono importanti per la diffusione del comportamento di voto nel mondo reale.

Complessivamente, gli autori hanno stimato che il "messaggio social" mostrato su *Facebook* ha aumentato l'affluenza diretta di circa 60.000 elettori e indirettamente, attraverso il contagio sociale, ha incrementato la partecipazione al voto di altri 280.000 elettori, per un totale di 340.000 voti aggiuntivi (corrispondenti allo 0,14% dell'elettorato).

L'esperimento di Bond et al. (2012) è stato in seguito replicato da Jones et al. (2017)[129] in occasione delle elezioni presidenziali del 2012 negli Stati Uniti. Nuovamente, si è osservato un significativo aumento di partecipazione al voto a seguito degli effetti diretti e indiretti dell'esposizione al "messaggio (*banner*) social". Anche in questo lavoro, inoltre, gli autori verificano un significativo aumento dell'affluenza tra gli amici intimi di coloro che hanno ricevuto il messaggio che incoraggiava a votare, e l'effetto totale sugli amici appare ancora più ampio dell'effetto diretto.

In definitiva, le evidenze rilevate in questo e nei precedenti paragrafi hanno dimostrato l'importanza dei *big data*, posti a fondamento dei meccanismi attraverso i quali operano le piattaforme online come i *social network*, nel sistema dell'informazione. Per mezzo delle piattaforme che li utilizzano, infatti, i *big data* giungono ad avere un effetto fondamentale sul pluralismo informativo, sia dal lato della domanda sia dal

[129] JONES J.J., BOND R.M., BAKSHY E., ECKLES D., FOWLER J.H., (2017), "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election", *PLoS ONE* 12(4).

lato dell'offerta (in termini tanto di risorse economiche estraibili dalla raccolta pubblicitaria quanto di personalizzazione algoritmica, selezione e prioritizzazione automatica dei contenuti informativi mostrati).

In tal senso, all'interno di un mezzo, internet – già di per sé utilizzato in maniera piuttosto ampia dagli individui politicamente più attivi, e quindi più schierati anche dal punto di vista ideologico, secondo dinamiche che portano alla formazione di *eco chamber* in cui le narrazioni possono arrivare ad essere autoreferenziali e le posizioni polarizzate[130] – l'operare congiunto delle azioni svolte sui *social network* dagli utenti più attivi e degli algoritmi di personalizzazione tende a favorire la diffusione virale della polarizzazione[131].

Più in generale, l'esposizione a messaggi informativi piuttosto che ad altri sulle piattaforme online, rispetto alla quale i *big data* raccolti risultano decisivi, non soltanto incide sulle percezioni degli utenti, ma è in grado di riflettersi sulla formazione delle opinioni degli stessi per poi tradursi in scelte e azioni concrete, incluse quelle determinanti per gli esiti elettorali.

3.4. L'approccio regolamentare dell'Autorità: il Tavolo Tecnico per la garanzia del pluralismo e della correttezza dell'informazione sulle piattaforme online

Da diverso tempo i fenomeni di disinformazione online, diffusi in special modo attraverso le piattaforme digitali di *big data*, e le ripercussioni degli stessi sull'opinione pubblica anche ai fini delle scelte politiche sono al centro di un percorso regolamentare, di analisi e di policy dell'Autorità, che risponde al cambio di paradigma imposto dall'affermarsi di una società (sempre più) *data-driven* (cfr. **Figura 3.10**).

In tal senso, l'Autorità, nell'ultimo quinquennio, ha condotto attività di monitoraggio e studio sulle specifiche modalità tecniche ed economiche di funzionamento delle piattaforme online come mezzi di accesso e distribuzione delle notizie oltre che sull'impatto di questi nuovi attori della rete sul sistema dell'informazione e sul pluralismo. L'Autorità ha così potuto attestare l'emergenza delle criticità connesse al crescente utilizzo di *social network* e motori di ricerca anche nelle campagne elettorali e referendarie, nonché la diffusione di strategie di disinformazione (basate tra l'altro sull'acquisizione di *big data*) mediante le piattaforme online.

L'evoluzione degli scenari informativi, tanto dal lato della domanda quanto da quello dell'offerta, è stata oggetto di ulteriori approfondimenti attraverso la stesura di Rapporti, l'organizzazione di *workshop* e lo svolgimento di Indagini conoscitive come quella avviata su "Piattaforme digitali e sistema dell'informazione" (delibera n. 309/16/CONS) e quella congiunta con l'Autorità garante della concorrenza e del mercato il Garante per la protezione dei dati personali in cui si inserisce il presente Rapporto.

Dalle analisi in corso è emerso, peraltro, che **l'esame dei fenomeni di disinformazione online richiede un approccio multidisciplinare, nonché l'adozione di iniziative di cooperazione e confronto con i soggetti operanti nel sistema dell'informazione online, le istituzioni di ricerca e le associazioni di settore, sia per acquisire una conoscenza adeguata di fenomeni complessi,**

[130] Per un'approfondita trattazione, cfr. AGCOM (2018), *Rapporto sul consumo di informazione*, cit.
[131] Cfr., tra gli altri, BESSI A., ZOLLO F., DEL VICARIO M., SCALA A., CALDARELLI G., QUATTROCIOCCHI W., (2015), "Trend of Narratives in the Age of Misinformation", *PLoS ONE* 10(8); QUATTROCIOCCHI W., A. Scala, C. R. SUNSTEIN (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110; QUATTROCIOCCHI W., VICINI A., (2016), *Misinformation: Guida alla società dell'informazione e della credulità*, FrancoAngeli; QUATTROCIOCCHI W., VICINI A., (2018), *Liberi di crederci: Informazione, internet e post-verità*, Codice edizioni.

come l'impatto delle piattaforme sull'opinione pubblica, sia per incoraggiare forme di autoregolamentazione degli operatori attivi nel sistema dell'informazione.

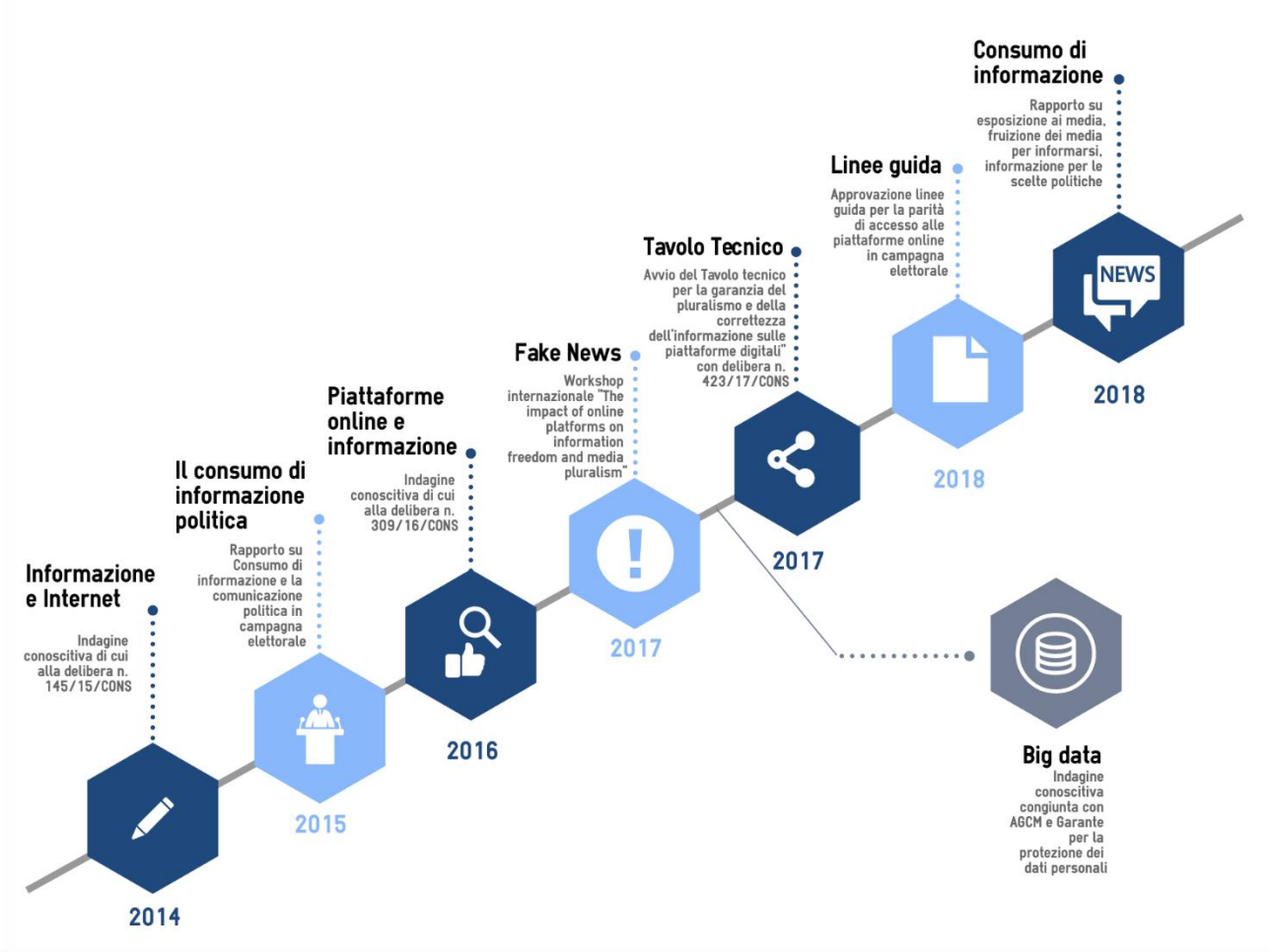


Figura 3.10 – Il percorso regolamentare di Agcom in materia di informazione online
Fonte: Autorità

Sulla base di queste premesse e in ragione dei mutamenti introdotti dall'utilizzo dei *big data*, l'Autorità, sul finire del 2017, ha istituito il "**Tavolo Tecnico per la garanzia del pluralismo e della correttezza dell'informazione sulle piattaforme digitali**" (delibera n. 423/17/CONS), con cui ha inteso perseguire l'obiettivo di promuovere l'autoregolamentazione delle piattaforme e lo scambio di buone prassi per l'individuazione e il contrasto dei fenomeni di disinformazione online frutto di strategie mirate. In particolare, lo scopo principale del Tavolo è favorire la condivisione di informazioni, il confronto e l'emersione di idonee metodologie di rilevazione, nonché l'individuazione degli strumenti di trasparenza, delle regole e tecniche di intervento più adeguate a garantire, specie nel corso delle campagne elettorali, parità di trattamento per tutti i soggetti politici presenti sulle piattaforme e correttezza e imparzialità dell'informazione per gli utenti.

L'esperienza del Tavolo promosso dall'Autorità rappresenta un *unicum* nel panorama mondiale. Sebbene l'adozione di strumenti analoghi di autoregolamentazione da parte delle piattaforme e degli editori, per esempio in materia di *fact-checking*, sia già stata sperimentata all'estero, **quello italiano è il primo caso di coordinamento degli attori del sistema informativo promosso da un'autorità indipendente che operi in funzione di facilitare il dialogo tra gli *stakeholder***. Nell'attuale fase di evoluzione del sistema,

l'Autorità ha ritenuto, infatti, di sostenere e monitorare le iniziative di autoregolamentazione poste in essere dalle imprese coinvolte, favorendo altresì il confronto e il contributo di esperti internazionali, Università, centri di ricerca e associazioni di categoria.

Nello specifico, come mostra la **Figura 3.11**, **al Tavolo Tecnico partecipano i rappresentanti di tutte le componenti del sistema informativo**: le piattaforme online; gli editori sia tradizionali con offerte informative online sia operanti esclusivamente in rete; i giornalisti; le associazioni del comparto pubblicitario e dei consumatori.

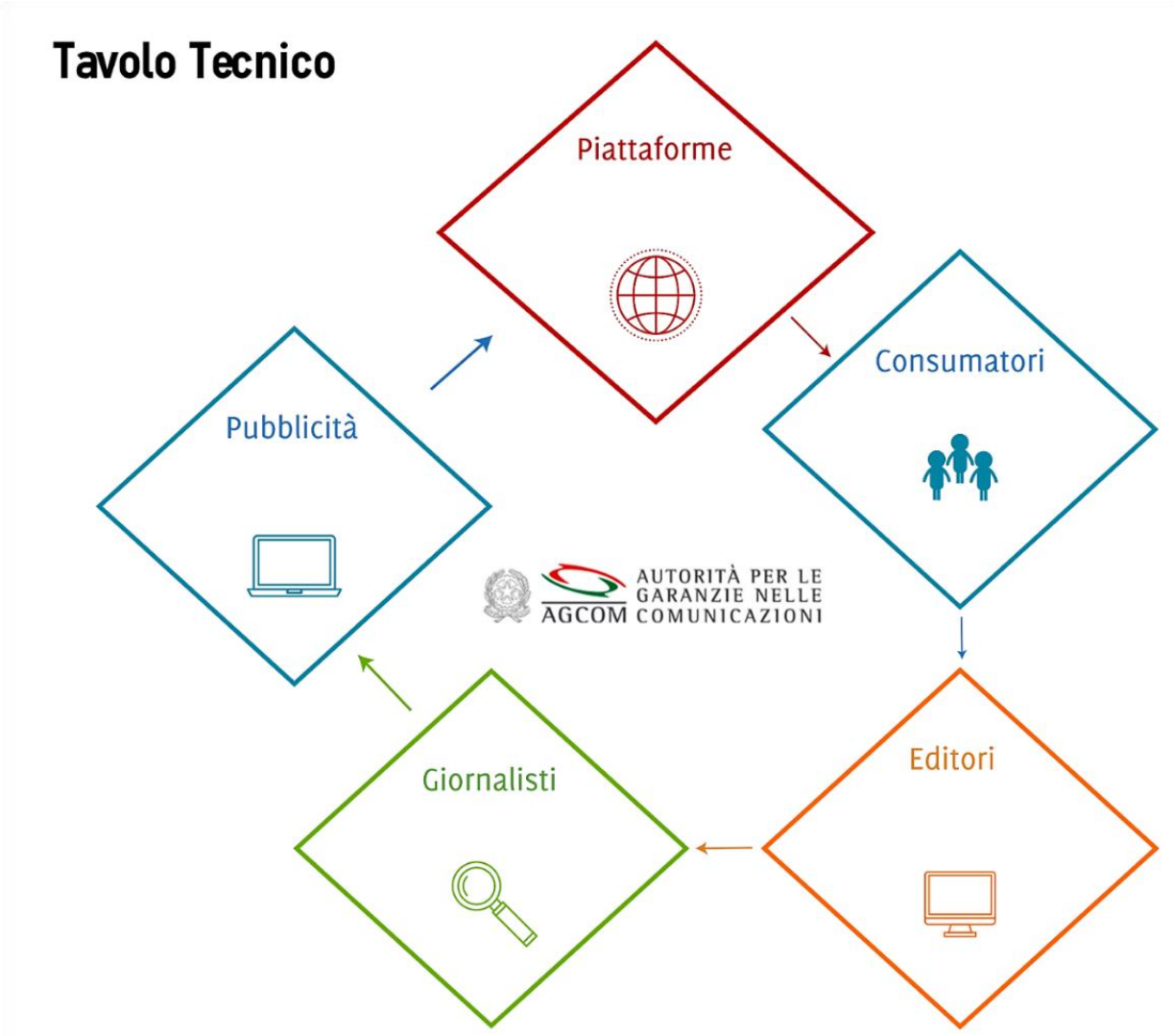


Figura 3.11 – I componenti del Tavolo Tecnico
Fonte: Autorità

Il Tavolo Tecnico, le cui attività sono organizzate in incontri, riunioni plenarie e gruppi di lavoro (v. **Figura 3.12**), ha avuto una prima fase operativa - corrispondente al periodo precedente l'avvio della campagna per le elezioni politiche del 4 marzo 2018 - volta all'individuazione, nel rispetto della libertà d'espressione, di strumenti di autoregolamentazione per:

i) la prevenzione e il contrasto di strategie di disinformazione online in campagna elettorale;

ii) la creazione di un dibattito libero e consapevole in rete, anche con riferimento a temi tipicamente oggetto di confronto politico-elettorale;

iii) la garanzia, anche nel contesto informativo contraddistinto dal ruolo delle piattaforme di *big data*, della parità di accesso per tutti i soggetti politici concorrenti in campagna elettorale.

In questo ambito, nel febbraio 2018, sono state adottate le "**Linee guida per la parità di accesso alle piattaforme online durante la campagna elettorale 2018**", grazie alle quali le piattaforme aderenti (*Google* e *Facebook*) hanno reso disponibili alcuni strumenti specifici, tra cui la campagna informativa lanciata da *Facebook* sulle pagine dei propri utenti italiani per l'individuazione delle notizie false, e servizi rivolti ai soggetti politici interessati a far conoscere il proprio programma ai cittadini (ad esempio, *Google Posts* e *Facebook Issues*). Nel contesto delle Linee guida, si inseriscono, inoltre, le attività di verifica e rimozione di contenuti illeciti, ovvero contrari alla normativa nazionale in materia di *par condicio* (ad esempio, sondaggi diffusi nei 15 giorni precedenti le elezioni), dietro segnalazione, e le esperienze di *fact-checking* portate avanti anche con il supporto di organizzazioni indipendenti.

Figura 3.12 – Le attività del Tavolo Tecnico
Fonte: Autorità

Alla prima fase operativa del Tavolo ne sta seguendo un'altra, che ha visto l'istituzione di cinque gruppi di lavoro inerenti a:

a) **metodologie di classificazione e rilevazione dei fenomeni di disinformazione online**;

b) **definizione dei sistemi di monitoraggio dei flussi economici** pubblicitari, da fonti nazionali ed estere, volti al finanziamento dei contenuti *fake*;

c) ***fact-checking***: organizzazione, tecniche, strumenti ed effetti;

d) ***media literacy*** e disinformazione online;

e) progettazione e realizzazione di **campagne informative su disinformazione rivolte ai consumatori**.

Nel corso delle attività dei gruppi di lavoro, è stato compiuto un preliminare sforzo definitorio per delineare in maniera condivisa gli aspetti peculiari e identificativi dei problemi legati alla disinformazione su internet. All'esito della ricognizione operata, **sono stati individuati gli elementi salienti da considerare per classificare i vari disturbi dell'informazione online (in mis-informazione, mala-informazione e disinformazione)**[132]. Si tratta di elementi che attengono alle fasi di produzione dei contenuti informativi (falsità dei contenuti; contagiosità degli stessi; intento doloso sottostante alla loro creazione; motivazione politico/ideologica o economica di chi li crea per poi diffonderli), diffusione degli stessi (in maniera massiva) e impatto per il pluralismo informativo (generazione di effetti sulla formazione dell'opinione pubblica).

Nell'individuare le prerogative dei contenuti suscettibili di costituire un disturbo informativo online e arrecare un pregiudizio al rispetto del principio pluralistico, è stato posto l'accento ancora una volta sull'importanza assunta dallo sfruttamento dei *big data* per la messa in atto di tecniche di diffusione massiva e propagazione virale di notizie false, attraverso le piattaforme online.

Proprio sul ruolo dei *big data* nella realizzazione di strategie di disinformazione in rete e, più in generale, sulla **ricostruzione della filiera dei falsi contenuti informativi online** si sta attualmente focalizzando l'analisi trasversale dei gruppi del Tavolo Tecnico, nella prospettiva di identificare le principali attività, le modalità organizzative, le tecnologie e le risorse utilizzate (compresi i *big data*) per la creazione, produzione e distribuzione dei contenuti *fake*, e la concreta attuazione delle strategie di disinformazione *online*. Tali strategie, infatti, sono caratterizzate dalla presenza di una struttura organizzata, che si pone obiettivi, di natura economica e non, di breve, medio e lungo periodo. In particolare, l'Autorità ha rinvenuto l'opportunità di effettuare un esame ad ampio spettro delle predette strategie, ricomprendendo sia quelle che si fondano su motivazioni di ordine economico, sia quelle che si basano su motivazioni ideologico-politiche. Dalle analisi finora effettuate, emerge un quadro complesso, in cui **coesiste una varietà di veri e propri modelli di *business* della disinformazione *online*, taluni basati sulla raccolta pubblicitaria, altri sul contributo diretto degli utenti mediante azioni fraudolente, altri ancora sulla diffusione di contenuti che mirano a danneggiare il marchio e l'immagine delle imprese che ne sono vittime**.

Nel prosieguo dei lavori del Tavolo saranno, dunque, indagati in dettaglio questi modelli, i flussi di risorse e le strategie non commerciali, anche con l'ausilio di *case study*, in modo da mettere in luce le criticità da affrontare per il contrasto della disinformazione online e definire la combinazione più adeguata di soluzioni tecniche, di mercato e codici di autoregolamentazione.

[132] In particolare, con il termine "mis-informatione" si identifica la categoria di contenuti informativi divulgati su internet non veritieri o riportati in modo inaccurato, suscettibili di essere recepiti come reali, ma non creati con un intento doloso. Per "mala-informazione" si intende la categoria di contenuti informativi fondati su fatti reali (anche a carattere privato) divulgati su internet e contestualizzati in modo da poter essere anche virali e veicolare un messaggio con il preciso intento di danneggiare una persona, un'organizzazione o un Paese, o affermare/screditare una tesi; mentre per "disinformazione" si intende la categoria di contenuti informativi, anche sponsorizzati, artatamente creati in modo da risultare verosimili, contraddistinti non solo dalla falsità dei fatti, ma anche dalla loro contagiosità, nonché dall'intento doloso di pubblicazione e diffusione in modo massivo.

Ciascun gruppo di lavoro del Tavolo, inoltre, sta proseguendo nelle attività specifiche afferenti alla propria tematica di riferimento, adottando un approccio analitico che, incardinandosi nel nuovo paradigma introdotto dalla diffusione pervasiva dei *big data* anche nell'ecosistema dell'informazione, non può che essere fondato sulla conoscenza e sulle più avanzate impostazioni metodologiche di tipo interdisciplinare e *data-driven*, ossia che presuppongono, esse stesse, l'impiego di grandi masse di dati.

In conclusione, il Tavolo Tecnico per la garanzia del pluralismo e della correttezza dell'informazione sulle piattaforme online costituisce un tassello fondamentale nel percorso conoscitivo e regolamentare in materia di informazione e disinformazione online che l'Autorità sta conducendo. **Il Tavolo Tecnico**, più precisamente, **è espressione di un nuovo approccio di analisi, regolamentazione e policy, orientato a conoscere per poter adeguatamente deliberare**. In altri termini, un approccio che trova fondamento nell'acquisizione, propedeutica e imprescindibile, di tutte le informazioni e conoscenze necessarie a investigare fenomeni che presentano un grado di complessità molto elevato, derivante non soltanto dal salto tecnologico connesso all'avvento dei *big data* e sottostante al funzionamento delle piattaforme online (per mezzo degli algoritmi), ma anche dall'estrema rilevanza dei molteplici diritti (individuali e collettivi) coinvolti, tra cui quelli sociali e politici che necessariamente richiedono misure tempestive ed *ex ante* per poter essere debitamente tutelati. **È attraverso esperienze come il Tavolo Tecnico, che l'Autorità – con tutti gli strumenti a disposizione, il confronto costante con le parti in causa e l'assidua collaborazione con il mondo scientifico internazionale – rende operativi i principi del nuovo approccio nel contesto regolamentare nazionale e comunitario di riferimento**.

Più in generale, a partire dalle evidenze sopra esposte, la fase successiva dell'Indagine conoscitiva sui *big data*, per gli aspetti pertinenza dell'Autorità, si concentrerà specificamente sulle tematiche relative al rapporto tra *big data* e pluralismo 2.0. In quella sede, inoltre, saranno elaborati possibili suggerimenti di policy e di evoluzione del quadro regolamentare.