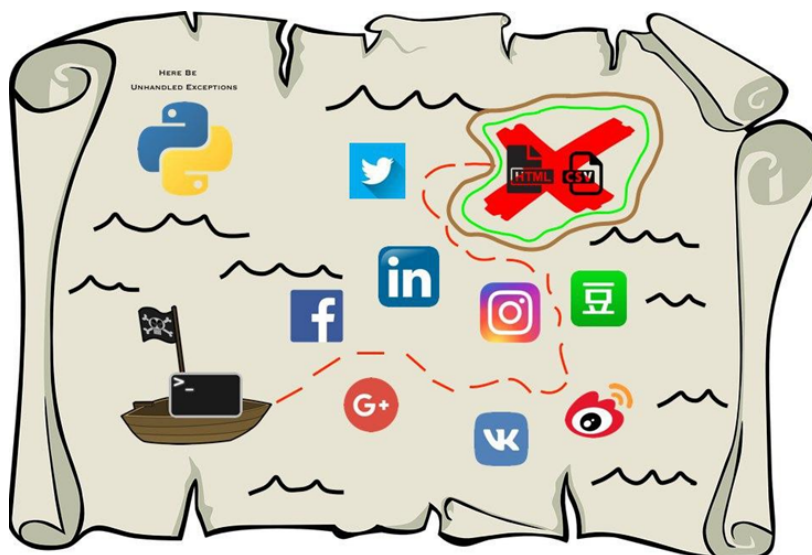




CORSO DI FONDAMENTI DI DATA SCIENCE & MACHINE LEARNING



Progetto: Social Mapper

PROFESSORE **Giuseppe Polese**

DOTTORE **Domenico Desiato**

DOTTORE **Stefano Cirillo**

IDEATORI

Francesca Cerruto Matricola: 0522500784

Michele Simone Gambardella Matricola: 0522500785

CURRICULUM DATA SCIENCE & MACHINE LEARNING

CAPITOLO 1

Introduzione

Oggigiorno con l'avvento della tecnologia chiunque dispone di un profilo web su almeno un social network. Molti sono i social presenti in rete ed ogni anno ne nascono di nuovi. Essi si aggiornano di continuo e rappresentano una vera e propria vetrina sulla vita delle persone tramite ciò che esse decidono di raccontare e di condividere con il mondo esterno. I social sono ormai utilizzati non solo per comunicare con altre persone ma anche a scopi pubblicitari e commerciali. L'elevata espansione dei social network ha comportato l'aumento sconsiderato della quantità di informazioni che ogni giorno i vari utenti condividono, molto spesso esponendo informazioni sensibili. Le varie aziende adottano, a tutela di tali dati, rigidi protocolli di sicurezza che spesso però si rivelano inutili poiché spesso è proprio l'utente a condividere tali informazioni, non curante dei rischi. Tali dati sensibili potrebbero essere sfruttati da malintenzionati per danneggiare gravemente altri utenti. La tutela della privacy sui social network è infatti un argomento poco toccato e a cui gli utenti danno spesso poco peso. Ogni social offre una ricca sezione dedicata alle politiche di privacy da adottare ma spesso queste vengono ignorate dagli utenti. La maggior parte delle informazioni di default sono visibili solo a persone appartenenti alla propria rete di contatti, ma questa politica può essere modificata permettendone la visione anche a persone esterne. La mole di dati sensibili condivisibili sui social ha creato l'esigenza di comprendere come gli utenti gestiscano la loro privacy e quante informazioni lascino trapelare. Da tali congetture nasce l'idea dello studio da noi svolto in questo progetto. Le nostre analisi sono state quindi svolte su persone sconosciute, esterne alla rete social degli account sfruttati per l'estrapolazione dei dati. Il nostro lavoro si incentra sulla costruzione di un dataset di circa 5000 utenti e sull'analisi delle informazioni condivise da tali persone, su tutti i social network a cui sono iscritti, per comprendere quali siano i dati più sensibili e quali quelli maggiormente condivisi.

Privacy Preserving on Social Network

Molteplici sono gli studi che sono stati effettuati in letteratura sul campo della privacy preserving. In questa sezione ci soffermeremo sull'analizzare alcuni lavori svolti a riguardo, i quali possono essere suddivisi in tre tipologie:

1. studi relativi a prevenire attacchi volti a violare la privacy degli utenti catturando informazioni sensibili;
2. studi relativi al mantenimento della privacy tra i diversi social network e dall'unione delle diverse informazioni;
3. studi e framework relativi alla decentralizzazione dei dati al fine di evitare la diffusione di informazioni sensibili a causa di attacchi mirati.

2.1 Prevenzione agli attacchi sui social

Per garantire la privacy sui social sono stati svolti vari studi e sono molte le tecniche che sono state messe in atto. Di seguito verranno trattati alcuni lavori svolti a migliorare la sicurezza su Facebook.

Tra i più significativi vi è uno studio dell'Università di Washington basato sui "Mi piace" messi dagli utenti. Questo studio ha evidenziato quanto anche un semplice "Like" sia un contenuto sensibile che può essere utilizzato dai social media e dal settore del marketing per carpire informazioni sugli interessi dell'utente e proporre pubblicità mirate. Questo studio ha fatto evincere quanto un semplice "Mi piace" possa aiutare a comprendere i dati demografici ed altri dati personali dell'utente. Con questo studio, il quale sfrutta l'algoritmo di classificazione KNN, essi sono riusciti a delineare, dai like messi dagli utenti, il loro sesso [1].

Lo studio di Adrienne Felt e David Evans dell'Università della Virginia evidenzia invece come l'introduzione di API aperte sui social network ha creato un modo per aggirare le impostazioni di controllo degli accessi e ha aumentato la facilità con cui gli attacchi possono essere eseguiti. Le applicazioni create utilizzando questa API

pongono seri problemi di privacy: un'applicazione "installata" acquisisce i privilegi del proprietario del profilo e può interrogare l'API per ottenere informazioni personali dell'utente o dei membri della sua rete. I dati a disposizione degli sviluppatori includono città natale, preferenze per gli appuntamenti e gusti musicali. Queste interfacce aperte sono sfruttate per i miglioramenti del sito, ma comportano gravi rischi per la privacy esponendo i dati degli utenti a sviluppatori di terze parti. Facebook ad esempio ha rilasciato la prima API di social networking per lo sviluppo di terze parti nel maggio del 2007. Il lavoro di Adrienne Felt e David Evans è quindi legato ad uno studio demografico su Facebook, svolto dagli studenti della CMU, il quale ha mostrato che l'89% degli utenti ha condiviso il sesso e la data di nascita completa alla propria rete social, mentre il 46% ha anche pubblicato la propria attuale residenza. Pertanto, le informazioni sulla maggior parte dei profili utente sono sufficienti per identificare in modo univoco i loro proprietari. Lo studio di questi due ricercatori permette di evitare questo problema, limitando l'acquisizione delle informazioni da parte della API, garantendo l'anonimato degli utenti grazie a dei proxy [2].

Uno studio delle Università delle Scienze e Tecnologie della Cina mira invece a prevenire un nuovo tipo di attacco, basato sulla violazione della privacy degli "utenti amici". Solitamente infatti gli utenti tendono a nascondere le proprie informazioni personali alle persone che non appartengono alla propria rete social, mentre sono soliti condividerle con gli "utenti amici". Un soggetto malintenzionato studiando la rete di amici di una persona può effettuare degli attacchi per violare la privacy di quest'ultimo. Per risolvere questo problema i ricercatori hanno implementato un nuovo algoritmo per garantire il mantenimento dell'anonimità degli amici in comune [3].

2.2 Social network collaborativi

Un'altra tipologia di studio significativa riguarda i social network collaborativi. Molti social condividono informazioni tra loro al fine di collegare i vari account. Tuttavia, senza un buon protocollo in grado di garantire la privacy, i dati sensibili degli utenti potrebbero diventare visibili agli altri senza che l'utente ne sia a conoscenza. Per risolvere questo problema Gary Blosser e Justin Zhan nel loro studio propongono come creare un social network collaborativo in grado di unire le informazioni delle varie piattaforme processando in maniera efficiente le query tra i vari social network. [4]

2.3 Decentralizzazione delle informazioni sui social

Gli studi di Leucio Antonio Cutillo e Refik Molva riguardano invece approcci utili per la decentralizzazione delle informazioni dei social network utili a prevenire la privacy degli utenti ed evitare attacchi da parte di utenti malintenzionati. Nei loro studi i ricercatori si fermano molteplici volte sul concetto della privacy sottolineando quanto la sensibilità delle informazioni condivise da parte degli utenti possa diventare materiale utile per carpire informazioni strettamente personali che possono essere utilizzate impropriamente [5], [6].

Un altro studio basato sulla decentralizzazione delle informazioni sui social, grazie alla costruzione di un apposito framework, è stato effettuato anche dall'Università di Duke [7].

2.4 Comparazione con lo stato dell'arte

Tra tutti gli studi esaminati, come si può notare, molti di essi si soffermano sul concetto di privacy, sui possibili attacchi che possono essere effettuati e su come prevenirli. Tuttavia, nessuno di essi si sofferma sull'effettuare un'analisi della privacy sui singoli social al fine di comprendere quali siano le informazioni più sensibili condivise dagli utenti sui differenti social network. Inoltre, nessuno degli studi precedentemente esaminati effettua un'analisi delle informazioni cross-social, al fine di evitare la ricostruzione di informazioni sensibili estrapolate dai differenti social network, ma ci si sofferma solo nell'unione delle informazioni da diversi social in un unico database al fine di effettuare interrogazioni più efficienti.

La creazione del nostro dataset, le analisi e le statistiche effettuate nel nostro progetto sono volte proprio a fornire queste informazioni al fine di sensibilizzare gli utenti sulla propria privacy e nel comunicare quali informazioni condivise possono essere altamente sensibili anche se considerate innocue da quest'ultimo.

In questo capitolo verranno illustrate le funzionalità iniziali di Social Mapper, le modifiche da noi effettuate e le nuove funzionalità implementate. Nell'ultima sezione viene poi esposta la procedura seguita per la creazione e il riempimento del dataset, utile per effettuare le analisi riportate nel Capitolo 4.

3.1 Funzionalità iniziali di Social Mapper

Social Mapper, implementato dal ricercatore Jacob Wilkin, permette di ricercare le persone su differenti social network quali: Facebook, LinkedIn, Instagram, VKontakte, Twitter, Pinterest, Weibo e Douban.

Il progetto è strutturato nelle seguenti componenti:

- **socialmapper.py:** Componente principale del progetto, in cui è definito il parser per l'immissione dei parametri da linea di comando e sono presenti le funzioni per ricercare le persone attraverso le varie modalità di input sui differenti social. Questa componente, infine, si occupa della formattazione dei risultati ottenuti e della costruzione del file di output, sia in formato HTML che in formato .csv;
- **un modulo differente per ogni social network:** come visibile in Figura 3.1 sono presenti 8 componenti differenziate che si occupano di effettuare il login sul sito web, di effettuare la ricerca del profilo e restituire il risultato in output.

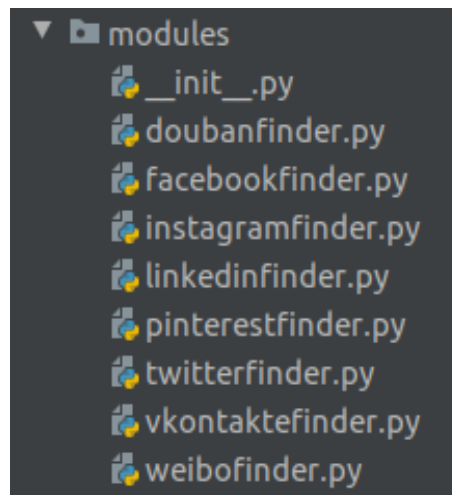


Figura 3.1: Moduli progetto Social Mapper

L'esecuzione può essere lanciata da linea di comando con diverse modalità di input:

- prendendo in input una cartella contenente le immagini delle persone da cercare denominate con il nome ed il cognome dell'individuo;
- tramite un file csv che ha come primo campo il nome ed il cognome dell'individuo e come secondo campo presenta il link ad un'immagine del soggetto da ricercare;
- tramite la ricerca delle persone per azienda su LinkedIn salvando tutte quelle trovate o in un'apposita cartella con il nome dell'azienda oppure in un csv inserendo il link dell'immagine trovata.

Social Mapper utilizza il riconoscimento facciale per individuare correttamente la persona sui diversi social network utilizzando le funzioni di libreria di *Face Recognition*. L'accuratezza nella ricerca può essere specificata tramite un parametro dato in input a linea di comando, utilizzando *"fast"* per ricerche più veloci ma meno accurate ed *"accurate"* per ricerche più accurate ma più lente.

Al termine delle ricerche Social Mapper restituisce due differenti output:

- un output in formato csv con il nome ed il cognome della persona da ricercare ed link a tutti i social network in cui l'individuo è stato trovato;
- un output in formato HTML con le foto prese in input con il nome e cognome nella prima colonna e le foto ritrovate sui differenti social network nelle altre colonne con i relativi link, come visibile in Figura 3.2.

Il progetto originale su GitHub, inizialmente scaricato, non funzionava su nessun social network poiché i siti web risultano in costante aggiornamento. Ciò comporta che le posizioni degli elementi e le classi dei *"div"*, dove risiedono le informazioni, cambino continuamente. In seguito alla natura mutevole delle pagine web, ciò comporta che i moduli di Social Mapper debbano essere costantemente aggiornati per rimanere in linea ai social network e per garantirne il corretto funzionamento.

Non è stato possibile però aggiornare i moduli dedicati ai social Weibo e Douban, poiché non ci è stato possibile creare un account fittizio per svolgere le nostre estrapolazioni.















Name	LinkedIn	Facebook	Twitter	Pinterest	Instagram
					
Bill Gates					
					
Steve Jobs					
					
					
Linus Torvalds					

Figura 3.2: Esempio output Social Mapper: pagina HTML

In tali social la registrazione richiede obbligatoriamente un numero di telefono per l'autenticazione, ma il suffisso italiano non risulta accettato. Non potendo quindi accedere alle informazioni dei profili su tali siti abbiamo preferito tralasciarli per concentrarci sui restanti social.

3.2 Funzionalità di crawling aggiunte

Come descritto Social Mapper inizialmente permetteva soltanto di ricercare le persone sui differenti social network, tuttavia per il nostro studio ciò non era sufficiente al fine di comprendere quanto gli utenti preservino la propria privacy nella diffusione delle informazioni sensibili. Risultava necessario poter estrarre dal profilo di ogni utente qualunque tipo di informazione esso mettesse a disposizione. Ad ogni modulo dei social network sono state dunque aggiunte delle funzionalità per navigare sul profilo di ogni persona correttamente individuata sul sito per andare ad estrarne tutte le informazioni rilevanti condivise. Oltre ad estrarre queste informazioni, per essere correttamente memorizzate nel csv, esse hanno subito le seguenti operazioni di pre-processing:

- creazione di un dizionario per ogni modulo formattato secondo i possibili campi estraibili dallo specifico social;
- i dati estratti sono stati ripuliti da spazi vuoti e valori anomali;
- i dati ripuliti sono stati interpretati e suddivisi nelle apposite sezioni del dizionario di cui facevano parte.

I dizionari ottenuti in output da ogni social sono stati utilizzati per memorizzare le informazioni all'interno del file csv fornito in output, che risulta contenere un campo per ogni possibile informazione presente nei dizionari.

3.3 Creazione Dataset

Sfruttando la funzionalità di Social Mapper di ricerca per il nome dell'azienda abbiamo raccolto circa 11000 foto di individui differenti, selezionando un elevato numero di nomi di aziende sia italiane che estere. I risultati ottenuti sono stati divisi in cartelle differenti ridenominate ognuna con il nome dell'azienda a cui fanno riferimento. Social Mapper è stato eseguito poi manualmente su ognuna delle cartelle ottenute, per estrapolare le informazioni condivise dai vari dipendenti sui propri profili. Per evitare di avere tuple

nel dataset che avessero tutti valori null è stato inserito il vincolo di memorizzare le informazioni relative all'utente, nel csv, soltanto se quest'ultimo era stato ritrovato su almeno un social network. Al termine di ogni esecuzione veniva restituito in output un differente file csv (per azienda) con tutte le informazioni estratte. Queste operazioni sono state svolte per ogni azienda ed hanno richiesto molto tempo a causa dell'elevato numero di persone da ricercare ed a causa della lentezza dei crawler che impiegano circa due ore per effettuare la ricerca in media di 150 individui. Tutti i csv raccolti sono stati successivamente integrati in unico file che è stato sottoposto ad ulteriori operazioni di ripulitura ed uniformazione dei dati effettuando le seguenti operazioni:

- eliminazione dei valori erroneamente estrapolati;
- eliminazione dei campi erroneamente suddivisi in colonne;
- uniformazione dei campi dei dati;
- pulizia dei dati da caratteri speciali (salvo biografie);
- inserimento di una colonna relativa all'azienda da cui è stato estratto ogni individuo.

Il dataset finale così costruito è composto da 4852 individui e contiene le seguenti informazioni:

- **Full Name:** il nome ed il cognome della persona cercata;
- **Linkedin:** il link al profilo Linkedin, se presente;
- **Cellulare:** il numero di cellulare della persona, se presente su Linkedin;
- **Sito Web:** sito personale o dell'azienda, se presente su Linkedin;
- **Email:** email personale, se presente su Linkedin;
- **Compleanno:** data di nascita, se presente su Linkedin;
- **Città:** città di residenza o di nascita, se presente su Linkedin;
- **Impiego:** posizione lavorativa, se presente su Linkedin;
- **Facebook:** link al profilo Facebook, se presente;
- **Work and Education:** posizione lavorativa ed informazioni relative all'istruzione scolastica, se presenti su Facebook;
- **Placed Lives:** luoghi in cui la persona è nata e vissuta, se presenti su Facebook;
- **Contact:** informazioni di contatto quali email, data di nascita, numeri di telefono se presenti su Facebook;
- **Basic Info Birthday:** data di nascita dell'individuo, se condivisa su Facebook;
- **Basic Info Gender:** sesso dell'individuo, se condiviso su Facebook;

- **Detail about:** biografia o informazioni personali dell'individuo, se presenti su Facebook;
- **Twitter:** link al profilo di Twitter, se presente;
- **Sito Twitter:** link al sito web dell'individuo, se presente su Twitter;
- **Città Twitter:** città di residenza o di nascita, se presente su Twitter;
- **Twitter Biografia:** biografia scritta dall'utente, se presente su Twitter;
- **Pinterest:** link al profilo Pinterest, se presente;
- **Instagram:** link al profilo Instagram, se presente;
- **Biografia Instagram:** biografia dell'utente, se presente su Instagram;
- **Vkontakte:** link al profilo VKontakte, se presente;
- **Data di Nascita:** data di nascita dell'utente, se presente su VKontakte;
- **Città1:** città di residenza, se presente su VKontakte;
- **Studiato a:** sede di studio, se presente su VKontakte;
- **Luogo di Nascita:** luogo in cui l'utente è nato, se presente su VKontakte;
- **Lingue:** lingue parlate dall'utente, se presenti su VKontakte;
- **Cellulare2:** numero di cellulare, se presente su VKontakte;
- **Telefono:** numero di telefono fisso, se presente su VKontakte;
- **Skype:** nickname Skype, se presente su VKontakte;
- **College o università:** nome del college o dell'università, se presente su VKontakte;
- **Stato:** stato professionale o livello massimo di istruzione, se presente su VKontakte;
- **Scuola:** scuole frequentate, se presenti su VKontakte;
- **Gruppi:** nomi gruppi a cui è iscritto l'utente, se presenti su VKontakte;
- **Azienda:** in cui l'individuo lavora, se presente su VKontakte;
- **Interesse:** interessi dell'utente, se presenti su VKontakte;
- **Azienda Estrapolata:** nome azienda a cui appartiene il dipendente.

Per le motivazioni sopra citate, ovviamente i dati riguardo i profili su Weibo e su Douban non sono presenti nel dataset, e non sono usati per le statistiche svolte nel Capitolo 4.

In questo capitolo sono mostrate le statistiche estrapolate a partire dai dati da noi ottenuti. Nella prima parte, ci siamo soffermati nell’analizzare le informazioni di ogni singolo social, cercando di comprendere pattern presenti che indicassero particolari caratteristiche sulla privacy mantenuta dagli utenti. Successivamente si è proceduto ad analizzare i vari dati incrociandoli tra loro cercando possibili relazioni tra i profili social, ricostruendo dati più complessi.

4.1 Statistiche per social

In questa sezione ci siamo soffermati sull’analisi dei singoli social, cercando di comprendere quali fossero le informazioni maggiormente condivise dagli utenti e quali fossero quelle più sensibili. Lo scopo di questa fase di analisi è stato capire se gli utenti rilasciassero dati sensibili sui vari social, senza inizialmente tenere conto di eventuali restrizioni o obblighi imposti dallo specifico social-media.

4.1.1 LinkedIn

Su LinkedIn sono state raccolte le seguenti informazioni riportate nella Tabella 4.1.

Informazioni LinkedIn	
Attributi	Numero di campi not-null
Cellulare	0
Sito Web	266
Email	59
Compleanno	141
Città	1441
Impiego	1447

Tabella 4.1: Attributi LinkedIn

Come si può notare in Figura 4.1 le informazioni maggiormente condivise dagli utenti sono state l'impiego lavorativo e la città. Tale valore fa riferimento al luogo di residenza o a quello di nascita, che può coincidere e per tanto non è stato escluso dall'analisi. Questo è dovuto dal fatto che LinkedIn è un social per il lavoro, creato per mettere in collegamento le aziende e i professionisti in cerca di lavoro. Gli utenti sono quindi maggiormente invogliati nel condividere questi dati. Pochi sono invece gli utenti che hanno condiviso anche le altre informazioni.

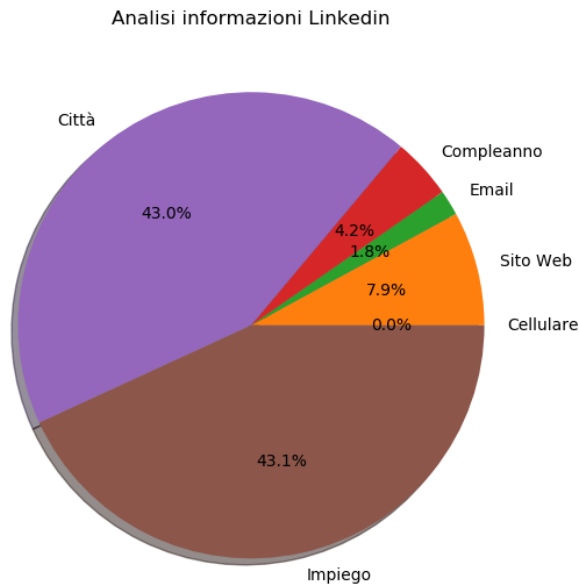


Figura 4.1: Statistiche dati LinkedIn

4.1.2 Facebook

Su Facebook sono state raccolte le seguenti informazioni riportate nella Tabella 4.2.

Informazioni Facebook	
Attributi	Numero di campi not-null
Work and Education	239
Placed Lives	255
Contact	0
Basic Info Birthday	0
Basic Info Gender	374
Detail About	0

Tabella 4.2: Attributi Facebook

Come si può notare in Figura 4.2 le informazioni maggiormente condivise dagli

utenti sono state quelle basilari (Basic info Gender) seguite dal luogo in cui essi vivono e dalle informazioni riguardanti la propria istruzione o il proprio ruolo lavorativo. Tra le informazioni basilari sono incluse a volte anche informazioni relative alla data di nascita (Basic info Birthday) degli utenti che, in combinazione con altre informazioni condivise, potrebbero intaccare notevolmente la loro privacy. Tra i soggetti presenti però nel nostro dataset tale eventualità non è presente poiché nessuno ha condiviso tale informazione.

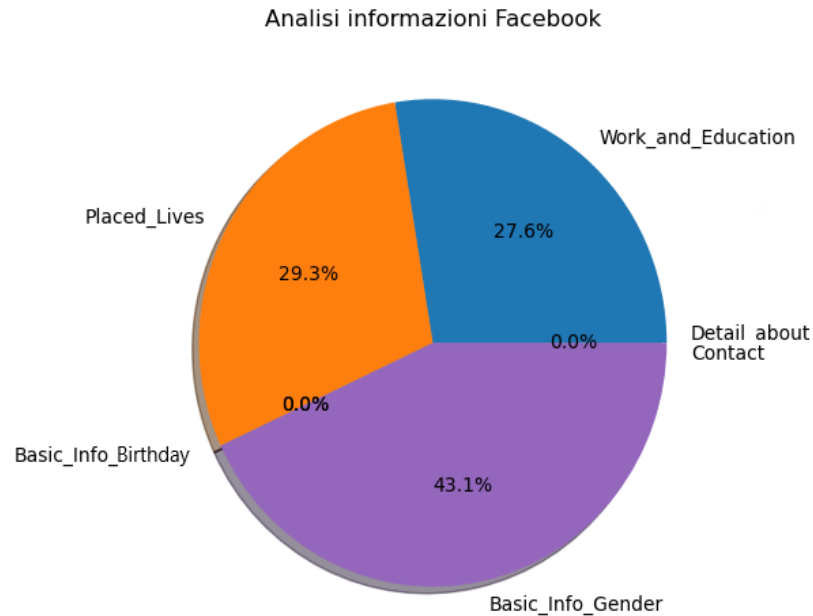


Figura 4.2: Statistiche dati Facebook

4.1.3 Twitter

Su Twitter sono state raccolte le seguenti informazioni riportate nella Tabella 4.3.

Informazioni Twitter	
Attributi	Numero di campi not-null
Sito Web	57
Città	78
Biografia	75

Tabella 4.3: Attributi Twitter

Poche solo le informazioni condivise su questo social network e pochi sono gli utenti del nostro dataset iscritti ad esso poiché maggiormente utilizzato da persone famose. Come si può notare dalla Figura 4.3 l'informazione maggiormente condivisa è relativa alla città d'appartenenza seguita dalla biografia e dal sito web. Nella biografia tuttavia un'utente può condividere qualsiasi informazione tra le quali anche un numero di telefono, un'email o altre informazioni sensibili.

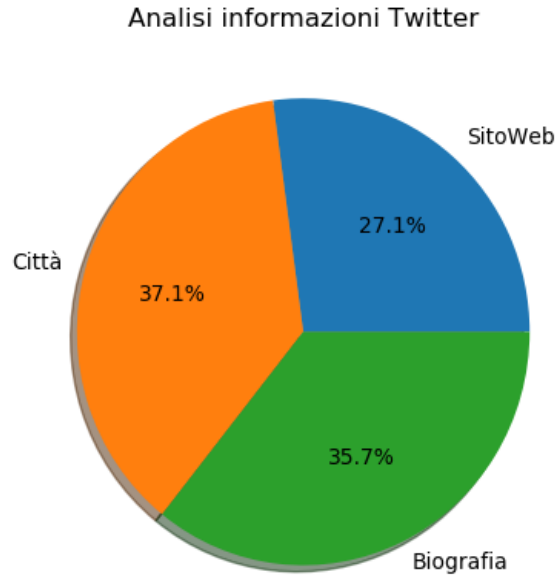


Figura 4.3: Statistiche dati Twitter

4.1.4 Instagram

Su questo social network gli utenti possono pubblicare principalmente video e foto pertanto l'unica informazione testuale che può essere estratta è relativa alla biografia dove un utente può scrivere qualunque cosa. Tuttavia, nemmeno quest'ultima può essere catturata se l'utente possiede un profilo privato. Data la natura molto variabile del contenuto informativo presente nelle biografie, tale dato non è stato trattato nelle analisi svolte.

4.1.5 VKontakte

VKontakte è un social non italiano simile a Facebook che permette agli utenti di condividere una mole enorme di informazioni tra le quali troviamo quelle presenti in Tabella 4.4.

Informazioni VKontakte	
Attributi	Numero di campi not-null
Data di Nascita	172
Città	47
Studiato a	54
Luogo di Nascita	38
Lingue	65
Cellulare	1
Telefono	1
Skype	26
College o Università	24
Stato	23

Scuola	49
Gruppi	70
Azienda	18
Interesse	1

Tabella 4.4: Attributi VKontakte

Come si può notare in Figura 4.4 l'informazione maggiormente condivisa dagli utenti del nostro dataset è stata la data di nascita seguita dai gruppi d'appartenenza e dalle lingue parlate. Poche, se non nulle, sono infatti le persone che hanno condiviso informazioni sensibili come il numero di telefono.

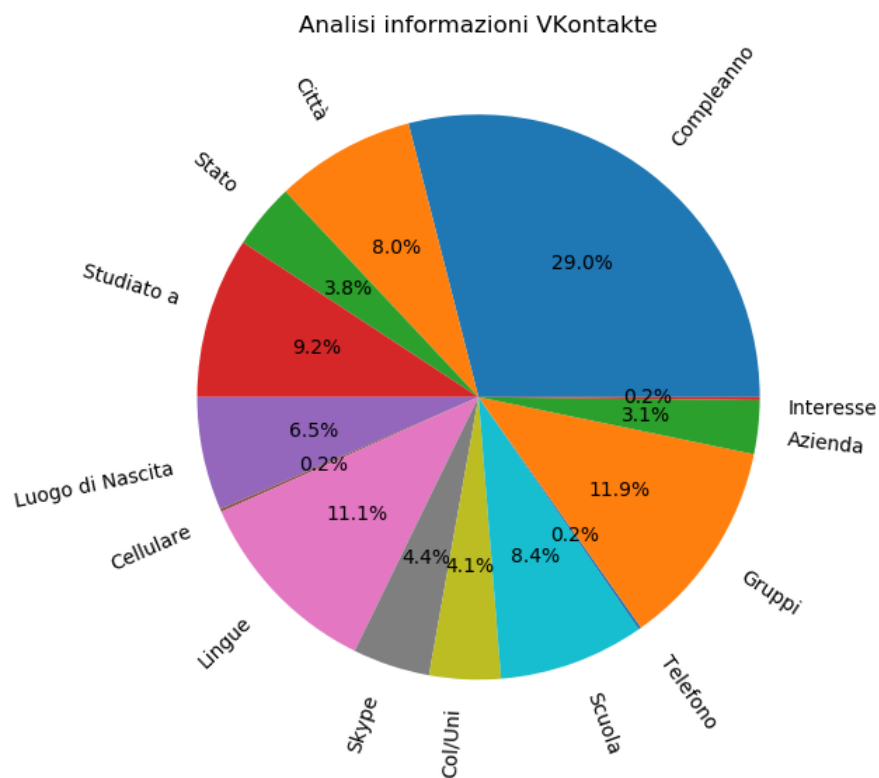


Figura 4.4: Statistiche dati VKontakte

4.2 Statistiche cross-social

A seguito delle nozioni acquisite nella precedente fase di analisi, abbiamo effettuato ulteriori statistiche, stavolta tenendo conto di tutti i social contemporaneamente per comprendere se alcune informazioni potessero essere ricostruite a partire da quelle rese pubbliche sui profili. Inizialmente abbiamo compreso la frequenza di condivisione delle varie informazioni presenti sui social per identificare possibili pattern, come visibile in Figura 4.5.

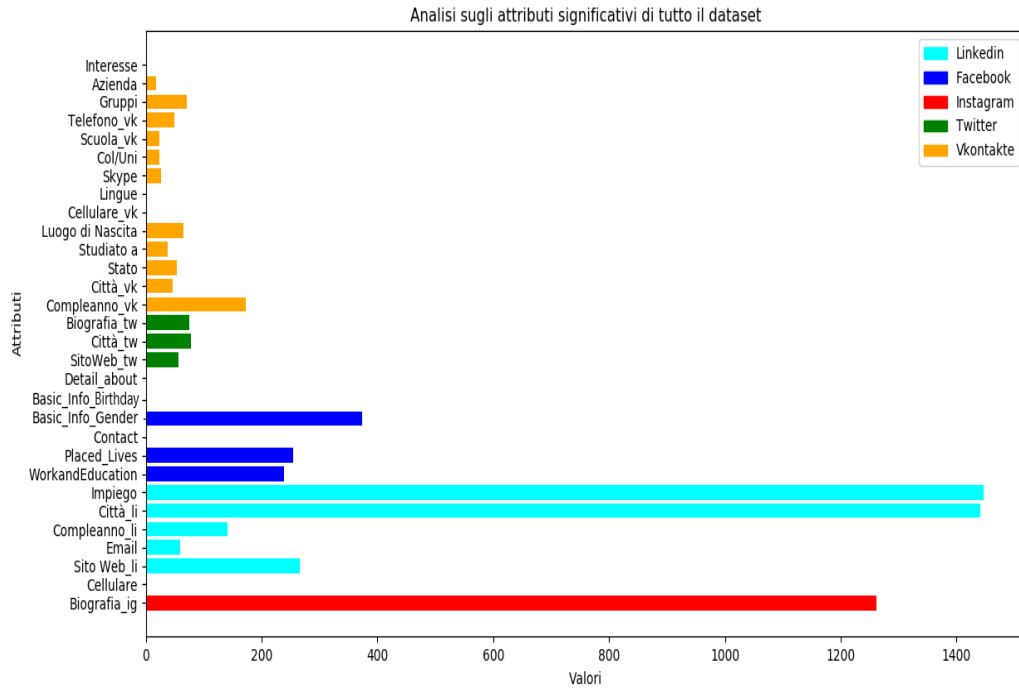


Figura 4.5: Statistiche attributi dell'intero dataset

Come visibile nel grafico la maggior parte degli utenti condividono molte informazioni su LinkedIn, segue poi la biografia su Instagram. Gli altri social presentavano una quantità di dati condivisi relativamente bassa. Tale differenza è dovuta principalmente alle restrizioni introdotte da LinkedIn durante la registrazione che richiede varie informazioni personali, le quali poi però non sono, nella maggior parte dei casi, privatizzate dagli utenti. Tale scelta però è dovuta dall'utilizzo prettamente a scopo lavorativo del social, per cui gli utenti cercano di condividere più dati possibili utili dal punto di vista professionale. Negli altri social invece, vi sono meno requisiti obbligatori e l'inserimento delle informazioni è a discrezione dell'utente.

4.2.1 Top-five

In Figura 4.6 è presente la top-five degli attributi maggiormente condivisi dagli utenti su tutti i social network. Al primo posto troviamo l'impiego seguito dalla città di residenza/di nascita, tali informazioni sono condivise su LinkedIn. Successivamente troviamo la biografia scritta dagli utenti su Instagram, le informazioni di base riguardanti il genere su Facebook ed il sito web personale condiviso su LinkedIn. Come si può notare il social su cui è pubblicata la maggiore quantità di informazioni è LinkedIn poiché, essendo un sito per la ricerca di lavoro, gli utenti sono maggiormente disposti a condividerle per poter, ad esempio, avere offerte lavorative o sponsorizzare la propria azienda.

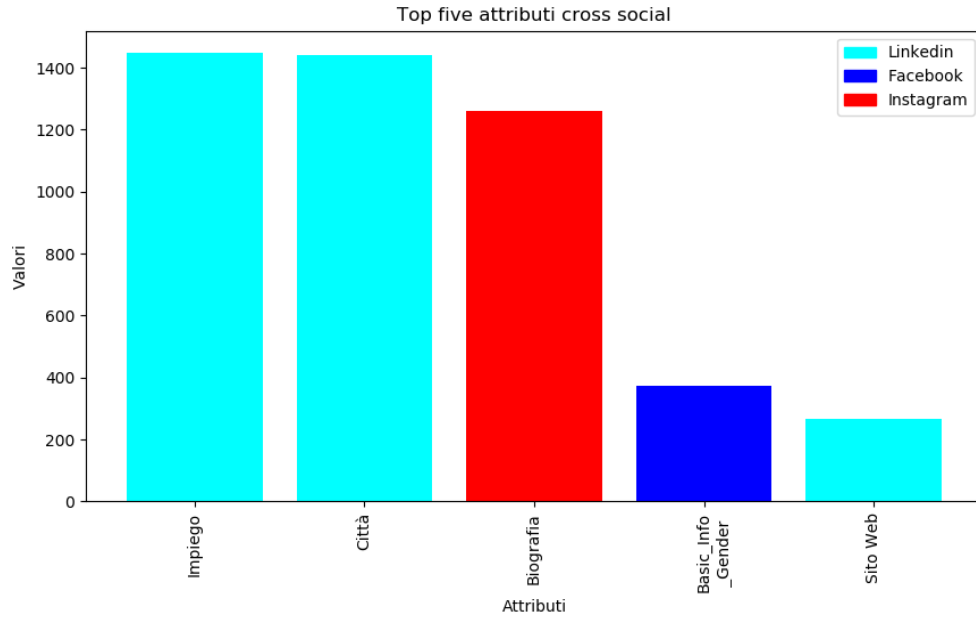


Figura 4.6: Top Five attributi maggiormente condivisi sui social

4.2.2 email cross-social

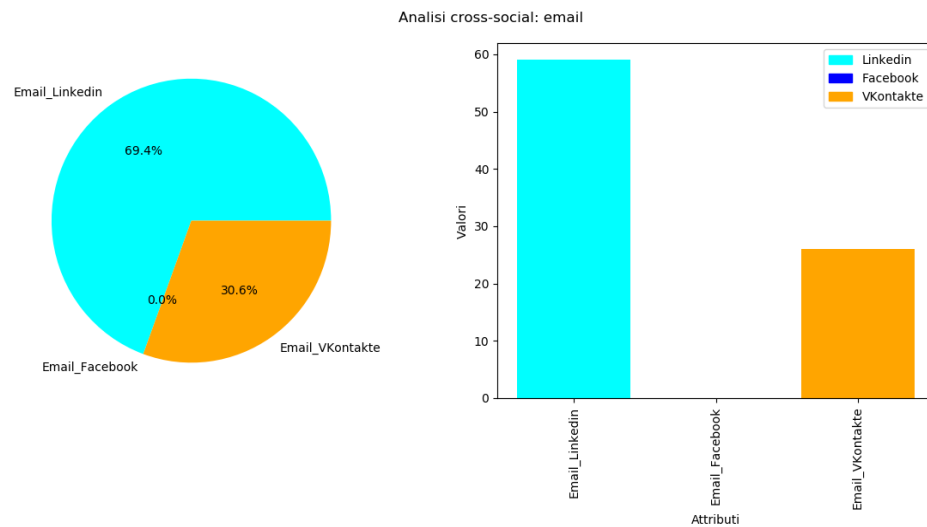


Figura 4.7: Statistiche sulle email

In Figura 4.7 sono presenti le statistiche riguardanti la condivisione dell'email. Analizzando i vari social è emerso che solo LinkedIn, Facebook e VKontakte presentano un'apposita sezione per l'inserimento specifico di questa informazione. A sinistra è riportato un grafico a torta che mostra il rapporto di distribuzione di tale dato. A destra, invece, vengono mostrate le frequenze assolute dell'email condivise per social. Visionando i grafici è facile comprendere che su LinkedIn gli utenti sono più propensi a condividere la propria email; tale social risulta prettamente usato da lavoratori che

quindi condividono informazioni di questo tipo per essere ricontattati da possibili datori di lavoro.

4.2.3 Data di nascita cross-social

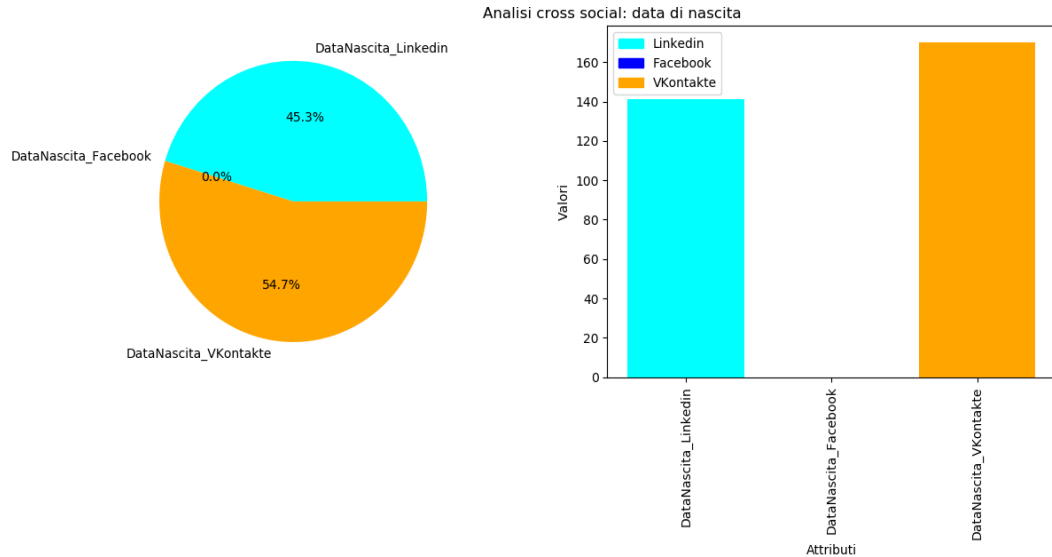


Figura 4.8: Statistiche sulle date di nascita

In Figura 4.8 sono presenti le statistiche riguardanti la condivisione della data di nascita. Analizzando i vari social è emerso che solo LinkedIn, Facebook e VKontakte presentano un'apposita sezione per l'inserimento specifico di questa informazione. A sinistra è riportato un grafico a torta che mostra il rapporto di distribuzione di tale dato. A destra, invece, vengono mostrate le frequenze assolute delle date di nascita condivise per social.

Visionando i grafici è possibile comprendere come la data del compleanno sia stata maggiormente condivisa su LinkedIn e VKontakte. Fornire la propria data di nascita su LinkedIn può essere un elemento discriminante utile per descrivere la carriera lavorativa dell'utente. Come si può notare su Facebook nessuno ha condiviso questa informazione poiché essa è contenuta nella sezione "Basic-Info" dove vi sono anche altri dati personali che la maggior parte degli utenti è propensa a condividere solo con account amici. In VKontakte la data di nascita è richiesta obbligatoriamente al momento dell'iscrizione e resa pubblica a tutti sul profilo. Una parte degli utenti del nostro dataset ha dunque però scelto di non privatizzare l'informazione.

4.2.4 Cellulare cross-social

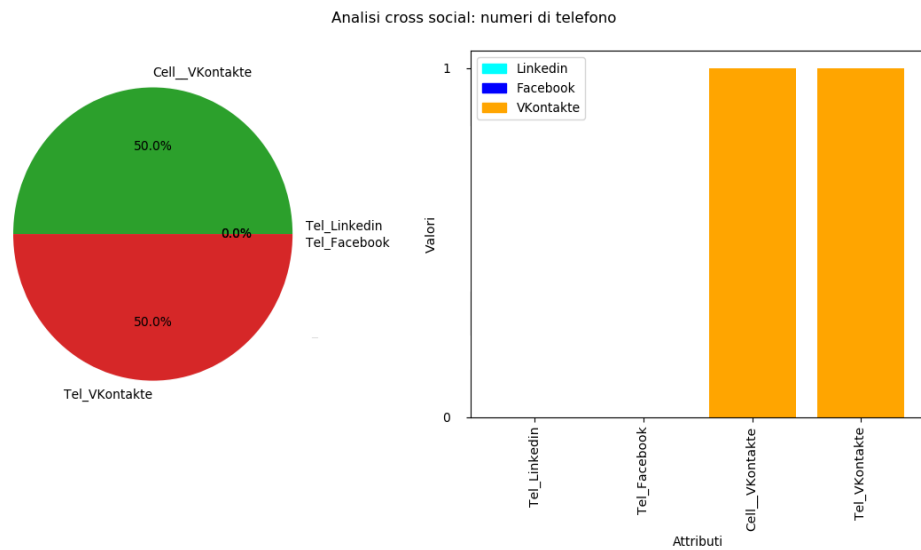


Figura 4.9: Statistiche sui numeri di Cellulare

In Figura 4.9 sono presenti le statistiche riguardanti la condivisione di un numero telefonico. Analizzando i vari social è emerso che solo LinkedIn, Facebook e VKontakte presentano un'apposita sezione per l'inserimento specifico di questa informazione. Nel dettaglio VKontakte permette di inserire sia un numero di telefono, sia uno di cellulare in un diverso campo. A sinistra è riportato un grafico a torta che mostra il rapporto di distribuzione di tale dato. A destra, invece, vengono mostrate le frequenze assolute dei numeri telefonici condivisi per social. Il numero di telefono permette di contattare chiunque in qualunque momento, per cui risulta essere un'informazione molto sensibile.

Da tale analisi è risultato che gli utenti gestiscano bene la privacy di tale informazione, tranne che su VKontakte. In tale social l'inserimento di un recapito telefonico è essenziale per la registrazione per cui ogni utente è tenuto a inserirlo. Successivamente è possibile poi nascondere tale dato andando a modificare le politiche del profilo, ma tale operazione, come si può vedere anche nei grafici, non è stata svolta da un solo utente.

4.2.5 Città cross-social

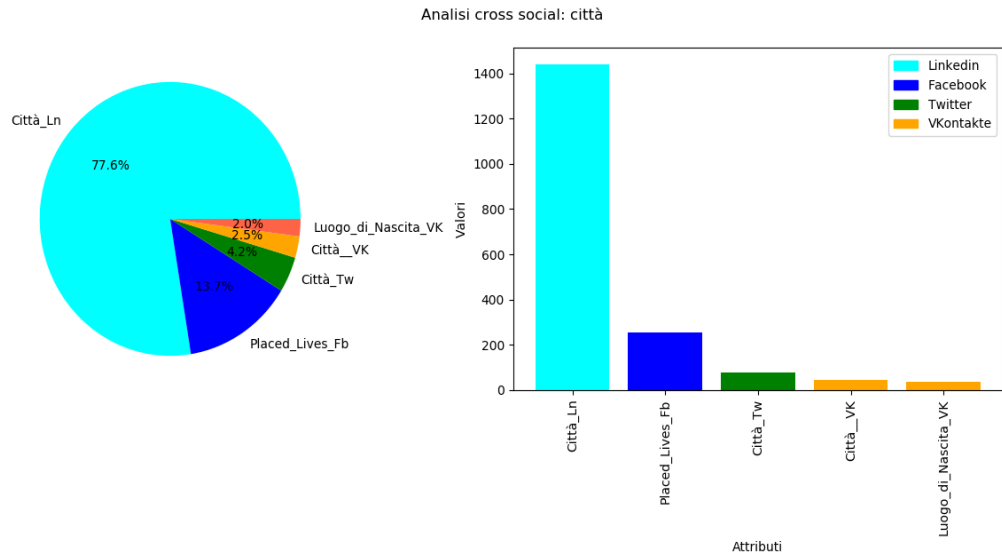


Figura 4.10: Statistiche sulle città

In Figura 4.10 sono presenti le statistiche riguardanti la condivisione della città di nascita o di provenienza. Analizzando i vari social è emerso che Twitter, LinkedIn, Facebook e VKontakte presentano un'apposita sezione per l'inserimento specifico di questa informazione. A sinistra è riportato un grafico a torta che mostra il rapporto di distribuzione di tale dato. A destra, invece, vengono mostrate le frequenze assolute delle città condivise per social. La città di nascita può essere considerata molto sensibile, poiché utile per la generazione del codice fiscale di una persona. Informazioni ulteriori circa altri luoghi possono comunque essere usate per ricostruire gli spostamenti di un soggetto.

Visionando i grafici è facile comprendere che su LinkedIn gli utenti sono più propensi a condividere tali dati; tale decisione può essere legata ai fini lavorativi per fornire informazioni a datori o ottenere proposte di impiego nelle vicinanze della propria abitazione. Spesso però la città mostrata risulta essere quella di residenza, più che quella natale, e quindi non utile agli scopi sopra elencati. Negli altri social invece sono di meno gli utenti che condividono tali dati. Su Facebook non sono presenti distinti campi che differenziano ad esempio la città di nascita da quella in cui ci si è trasferiti, per questo non è certo che le informazioni estrapolate contengano realmente la città natale.

4.2.6 Formazione o Lavoro cross-social

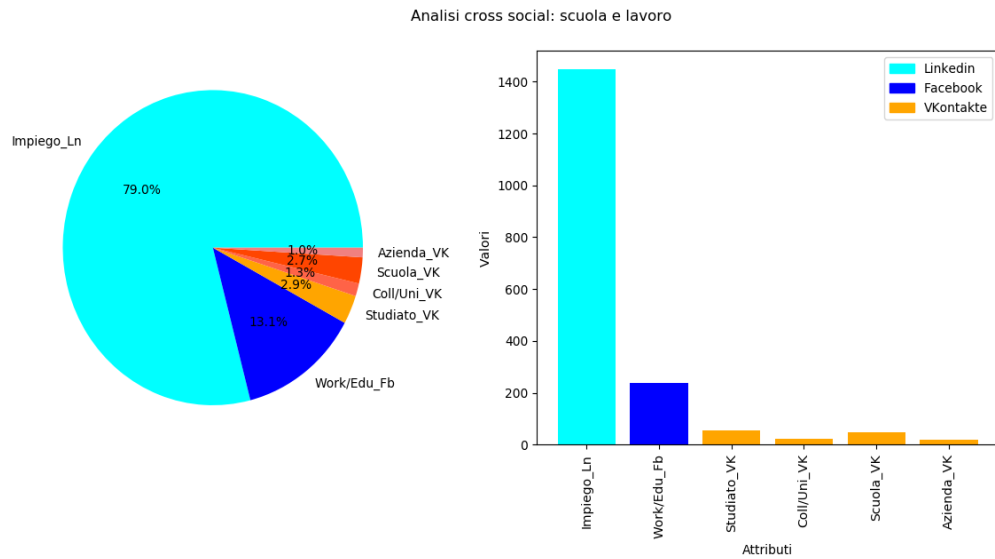


Figura 4.11: Statistiche sulle informazioni circa formazione o situazione lavorativa

In Figura 4.11 sono presenti le statistiche riguardanti le informazioni circa la formazione e gli impieghi lavorativi. Analizzando i vari social è emerso che solo LinkedIn, Facebook e VKontakte presentano un'apposita sezione per l'inserimento specifico di questa informazione. A sinistra è riportato un grafico a torta che mostra il rapporto di distribuzione di tale dato. A destra, invece, vengono mostrate le frequenze assolute delle varie informazioni condivise per social.

Visionando i grafici è visibile come tali dati siano condivisi maggiormente su LinkedIn, social usato prettamente ai fini lavorativi. In questo caso è molto importante esporre quante più informazioni personali quali skill possedute o traguardi di studio e/o lavorativi, al fine di enfatizzare le proprie competenze ed avere più proposte di lavoro possibili. LinkedIn non vincola la quantità di informazioni da poter condividere proprio per non imporre limiti agli utenti nella creazione di un profilo più o meno ricco. VKontakte, invece, ha una sezione molto strutturata per tali dati, sono presenti molte voci differenti riguardanti i vari lavori svolti e i risultati accademici conseguiti, i quali sono suddivisi per livello di formazione.

4.2.7 Analisi dati sensibili: Numero di cellulare

Uno dei dati più sensibili che può essere condiviso sui social è il numero di cellulare. Nella Sottosezione 4.2.4 abbiamo semplicemente analizzato la distribuzione delle informazioni condivise circa i numeri telefonici. Di seguito abbiamo approfondito lo studio andando a lavorare su tutto il dataset e verificando se la stessa persona abbia o meno condiviso tale dato su uno o più social. Quasi tutti i social, ad eccezione di Instagram, permettono di inserirlo. VKontakte ad esempio permette di registrarsi soltanto grazie al numero di cellulare che diventa il nickname univoco dell'utente. Questo dato diventa visibile a tutti gli utenti se non nascosto in un secondo momento. Dalla Figura 4.12 si evince che soltanto l'0.01% delle persone ha mantenuto il proprio numero di telefono pubblico sui social, in dettaglio soltanto su VKontakte. Il restante 99.99% non ha condiviso il proprio numero di telefono, andando dove necessario a modificare le politiche di privacy per non renderlo pubblico.

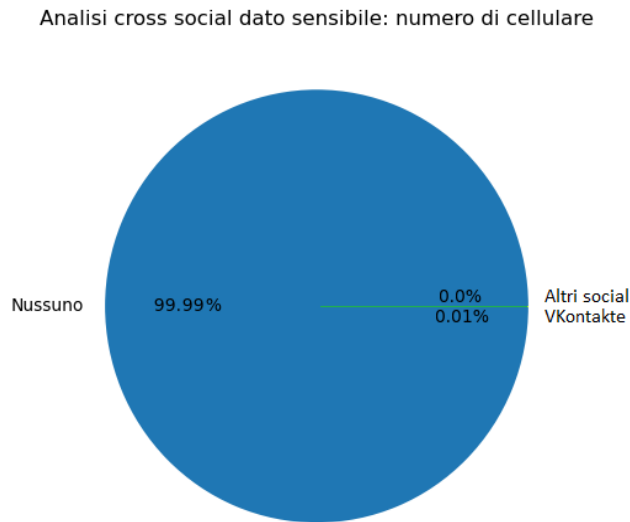


Figura 4.12: Statistiche sull'estrapolazione cross-social del numero di telefono

4.2.8 Analisi cross-social finali: Ricostruzione codice fiscale

Alcune informazioni condivisibili sui social sembrano banali e di poca importanza, tuttavia se messe insieme possono permettere di ricostruire informazioni sensibili dell'utente come il codice fiscale. Esistono infatti delle vere e proprie applicazioni disponibili anche su internet che permettono di ricostruirlo prendendo in input delle semplici informazioni quali la città natale, la data di nascita, il sesso, il nome ed il cognome dell'individuo. Abbiamo dunque effettuato un'analisi per scoprire quante persone abbiano condiviso queste informazioni su un singolo social oppure se divise tra differenti social. Ad esempio, una persona potrebbe decidere di condividere la data di nascita solo su LinkedIn e la città natale solo su Facebook ma unendo le informazioni tra i differenti social, incrociando i profili, questo ci permette di ricostruirne il codice fiscale. Come si può vedere in Figura 4.13 sono state effettuate queste analisi solo sul dato città natale e data di nascita poiché tutti gli utenti del nostro dataset possiedono il nome ed il

cognome, mentre il sesso può essere stato condiviso su Facebook. In assenza del genere è però facilmente deducibile dal nome dell'utente. Per motivi pratici le combinazioni di due o più siti sono state rappresentate tramite le iniziali dei vari social.

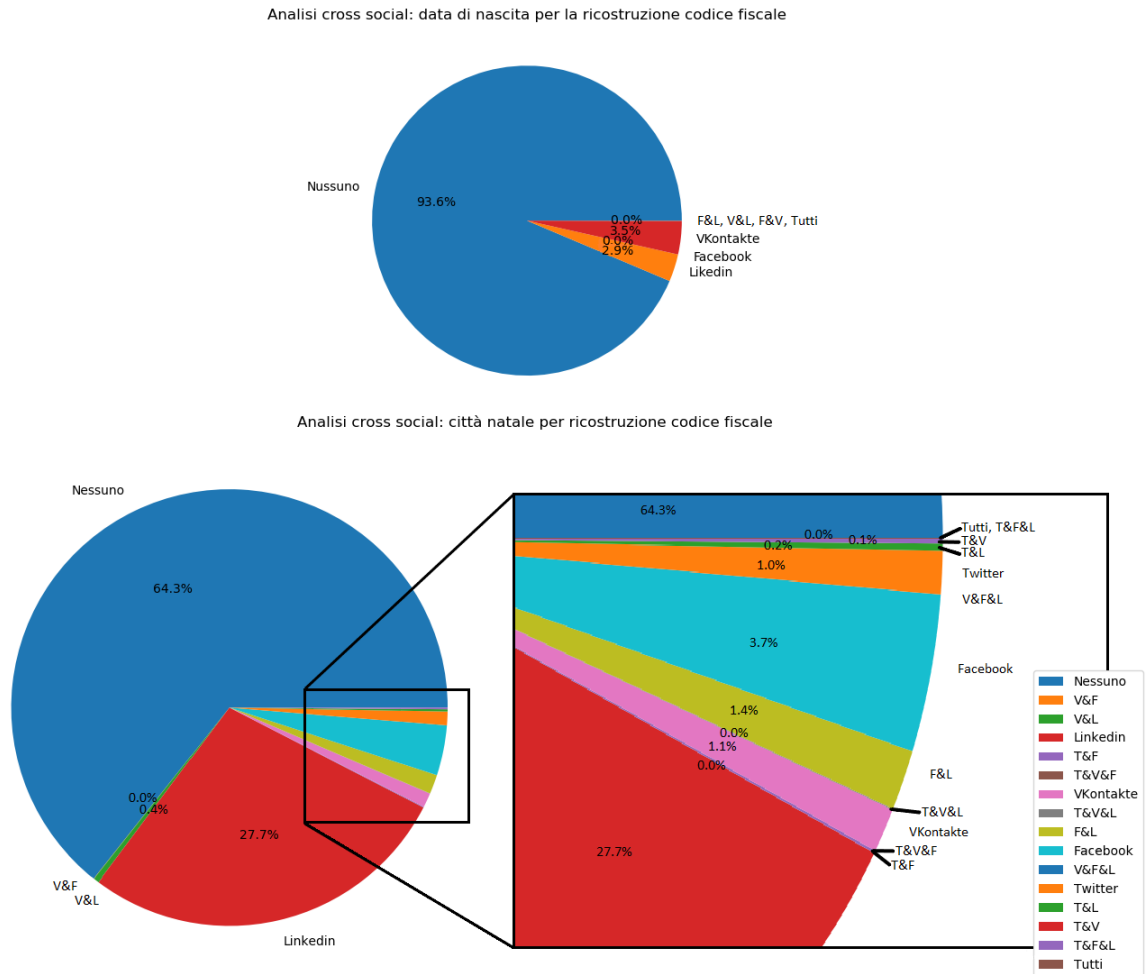


Figura 4.13: Statistiche sull'estrapolazione cross-social della data di nascita e della città natale

Nel grafo riguardante la data di nascita è facile notare come solo il 6.4% delle persone abbia condiviso tale dato in rete. Nel dettaglio gli utenti hanno reso pubblica la data in al massimo un social, infatti non ci sono casi dove tale informazione sia presente in due o in tutti i siti che lo permettono. Rispetto al secondo grafico invece possiamo comprendere come molte più persone abbiano condiviso in rete il luogo della propria nascita. Principalmente tale dato è reso pubblico su LinkedIn e Facebook. Il motivo per cui circa il 28% degli utenti abbia condiviso il dato su LinkedIn può essere giustificato dal fatto di poter voler essere contattati maggiormente da datori in una certa zona, a tal proposito molti aggiungono anche informazioni circa il luogo di residenza o dell'attuale lavoro.

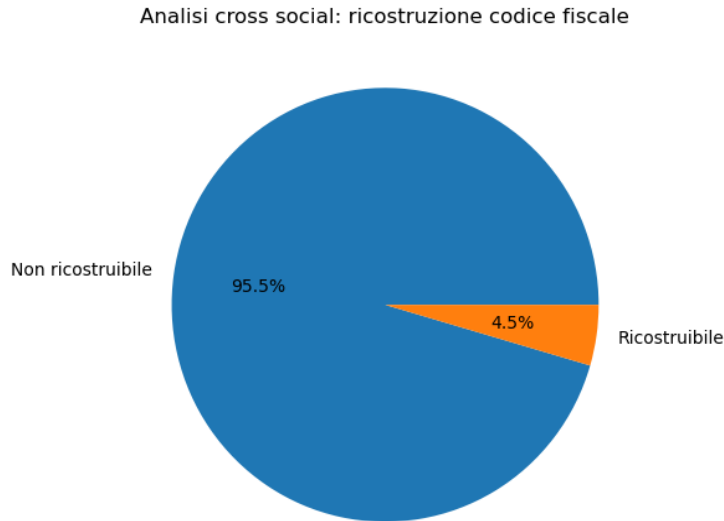


Figura 4.14: Statistiche sulla ricostruzione del codice fiscale tramite dati estrapolati

Il grafico in Figura 4.14 invece mostra la percentuale di persone del nostro dataset che ci permette di ricostruire il proprio codice fiscale. Come si può osservare la percentuale è del 4.5%, ciò quindi comporta che una parte del 6.4% delle persone che aveva condiviso la data di nascita non ha reso pubblico la città natale. La percentuale di codici ricostruibili, anche se può sembrare piccolissima considerando le circa 5000 persone del nostro dataset, corrisponde a circa 300 persone, numero che diventa fortemente significativo considerando la sensibilità del dato.

Condividere informazioni personali può essere molto pericoloso ed è necessario conoscere quali siano le informazioni sensibili da non condividere con leggerezza. L'obiettivo e monito di questo lavoro è stato quello di sensibilizzare gli utenti nella pubblicazioni di dati personali in tutti i social, e non solo su alcuni.

Il lavoro di Jacob Wilkin su Social Mapper permetteva semplicemente di ricercare in diversi modi le persone sui differenti social. Il nostro lavoro iniziale è stato quello di ampliarne le funzionalità permettendo di estrarre anche ogni possibile informazione condivisa dagli utenti sui social network. Il dataset così ottenuto ci ha permesso di realizzare numerose analisi al fine di comprendere quali siano le informazioni maggiormente condivise dagli utenti sulle diverse piattaforme e come le persone gestiscano la propria privacy sui singoli social e tra i vari social network. Come anche precedentemente illustrato un'utente può decidere di memorizzare un'informazione su un social, poiché ritenuto necessario, e non condividere quello stesso dato su di un altro privatizzandolo. Come dimostrato con le analisi da noi effettuate sul nostro dataset, estrapolando le informazioni degli utenti da tutti i social network questa privatizzazione del dato viene persa poiché un'informazione non disponibile su Facebook può essere reperita ad esempio su LinkedIn. Con tali strumenti è possibile ricostruire dati molto sensibili che minano la privacy degli utenti. Come illustrato allo stato dell'arte non è stato effettuato nessun lavoro d'analisi di questo tipo ma solo studi sui possibili attacchi alla privacy degli utenti. Al termine di tutte le analisi effettuate possiamo dedurre che il social su cui sono state condivise la maggior parte delle informazioni è LinkedIn, ritenuto altamente utile nel mondo del lavoro, seguito dalla biografia di Instagram e da Facebook. Dati fortemente sensibili, come il numero di cellulare, non sono stati condivisi da nessuno ad eccezione di un unico utente. Ciò fa comprendere che le persone analizzate siano a conoscenza dell'importanza della privacy dei propri dati sui singoli social network. Tuttavia, effettuando le analisi cross-social abbiamo compreso come gli utenti si preoccupino in minor modo di limitare la ricostruzione di dati importanti tramite l'incrocio di più profili su differenti social.

In conclusione, dal nostro studio possiamo dire di aver identificato le informazioni sensibili maggiormente condivise dagli utenti in rete. Abbiamo compreso l'importanza della gestione delle politiche di privacy non solo sul singolo social ma anche tra i differenti social, tematica molto delicata su cui la maggior parte degli utenti deve essere sensibilizzata.

Bibliografia

- [1] S. Bhagat, K. Saminathan, A. Agarwal, R. Dowsley, M. De Cock, and A. Nascimento, “Privacy-preserving user profiling with facebook likes,” 2018.
- [2] A. Felt and D. Evans, “Privacy protection for social networking platforms,” 2008.
- [3] C. Sun, S. Y. Philip, X. Kong, and Y. Fu, “Privacy preserving social network publication against mutual friend attacks.”
- [4] G. Blosser and J. Zhan, “Privacy preserving collaborative social network,” 2008.
- [5] L. A. Cutillo, R. Molva, and T. Strufe, “Privacy preserving social networking through decentralization,” 2009.
- [6] L. A. Cutillo, R. Molva, and T. Strufe, “Safebook: A privacy-preserving online social network leveraging on real-life trust,” *IEEE Communications Magazine*, vol. 47, no. 12, pp. 94–101, 2009.
- [7] A. Shakimov, H. Lim, R. Cáceres, L. P. Cox, K. Li, D. Liu, and A. Varshavsky, “Vis-a-vis: Privacy-preserving online social networking via virtual individual servers,” 2011.