Anonimizzazione di Dataset e Impatto sullo Schema Matching: Una Valutazione con ARX, Valentine e Automazione LLM

Zanni Davide
Università degli studi di Modena e Reggio Emilia
Dipartimento di Ingegneria "Enzo Ferrari"
250960@studenti.unimore.it

***Abstract*—**
L'anonimizzazione dei dati personali costituisce una sfida cruciale nel contesto della data science, in cui l'equilibrio tra tutela della privacy e mantenimento dell'utilità informativa rappresenta un obiettivo complesso da raggiungere. Questo lavoro propone una pipeline integrata che combina l'impiego di Large Language Models (LLM) per la generazione automatica di gerarchie di generalizzazione con l'utilizzo del framework ARX per l'applicazione di tecniche di k-anonymity su dataset tabellari.

L'approccio si fonda su un'automazione intelligente del processo di costruzione delle gerarchie, basata su analisi semantiche condotte dal LLM e successiva normalizzazione strutturale, riducendo la dipendenza dall'intervento manuale e migliorando la coerenza delle rappresentazioni.

La pipeline è articolata in quattro fasi principali: pre-processing e pulizia dei dati, generazione della gerarchia tramite LLM, validazione automatica delle strutture prodotte e anonimizzazione con ARX.

Gli esperimenti condotti su dataset sintetici — arricchiti con rumore controllato e attributi di diversa tipologia — mostrano l'impatto del livello di anonimizzazione (k = 2, 5, 10) sulle prestazioni di schema matching, evidenziando un chiaro trade-off tra protezione della privacy e degradazione delle metriche di accuratezza.

I risultati sottolineano la potenzialità dell'integrazione tra tecniche di apprendimento linguistico e strumenti classici di anonimizzazione, pur mettendo in luce le sfide legate alla stabilità semantica delle gerarchie generate e alla perdita informativa derivante dai processi di generalizzazione.

I. Introduzione

La diffusione pervasiva di dati personali nell'era digitale contemporanea ha reso la tutela della privacy una delle sfide centrali e prioritarie nella gestione, condivisione e analisi dell'informazione su larga scala [1]. Organizzazioni pubbliche, imprese e istituti di ricerca devono spesso pubblicare o integrare dataset contenenti informazioni sensibili (per esempio dati sanitari, finanziari o demografici) bilanciando l'utilità statistica con la protezione dell'identità dei soggetti.

A. Il Problema della Re-identificazione

La semplice rimozione degli identificatori diretti (ad esempio, nome, codice fiscale, indirizzo email) si è rivelata insufficiente a garantire l'anonimato, poiché combinazioni di attributi apparentemente innocui, denominati quasi-identificatori (QI) (data di nascita, genere o codice postale) possono consentire la re-identificazione di individui tramite correlazione con fonti esterne di dati. Ricerche empiriche hanno documentato che fino all'87% dei soggetti europei può essere univocamente identificato mediante l'incrocio di soli tre attributi quasi-identificativi con dataset pubblicamente disponibili.

B. Quadro Normativo: GDPR e Conformità

A livello normativo, il Regolamento Generale sulla Protezione dei Dati (GDPR, General Data Protection Regulation) – entrato in vigore nel 2018 nell'Unione Europea – ha stabilito principi e obblighi stringenti sulla gestione dei dati personali, promuovendo pratiche quali la pseudonimizzazione e l'anonimizzazione come strumenti cardine per la conformità legale [1].

C. K-Anonymity come Modello Formale di Protezione

Tra le metodologie consolidate per la protezione della privacy, il modello di *k-anonymity* rappresenta un paradigma rigoroso e fondamentale nella letteratura della privacy-preserving data publishing. Un dataset è detto *k-anonimo* se ogni combinazione di quasi-identificatori appare in almeno k record distinti, rendendo ogni individuo indistinguibile da almeno altri $k-1$ soggetti all'interno del dataset.

Questa proprietà formale viene ottenuta applicando trasformazioni controllate sui dati originari, tipicamente generalizzazione e soppressione, che riducono la specificità delle informazioni, assicurando privacy ma introducendo inevitabilmente una perdita di utilità informativa. In pratica, la scelta dei parametri di anonimizzazione riflette un compromesso (trade-off) fondamentale tra protezione della privacy e mantenimento della qualità analitica e della rappresentatività dei dati.

Sebbene la *k-anonymity* sia apprezzata per la sua semplicità concettuale e la sua efficienza computazionale, essa presenta alcune limitazioni strutturali, tra cui la vulnerabilità a attacchi di omogeneità (homogeneity attacks) e quelli basati su conoscenza esterna (background knowledge attacks), che possono compromettere la riservatezza anche in dataset formalmente conformi al criterio di anonimato.

D. Integrazione tra Anonimizzazione e Schema Matching

Uno degli aspetti critici e poco esplorati nell'uso di dati anonimizzati riguarda la qualità delle operazioni successive

di schema matching – ossia il processo di individuazione automatica delle corrispondenze semantiche tra colonne di dataset differenti. Tale operazione è fondamentale per attività centrali di data integration, dataset discovery e costruzione di data lake, ma risulta particolarmente sensibile alle modifiche introdotte dall'anonimizzazione, poiché le distribuzioni di valori, i domini e la granularità semantica vengono alterati durante il processo di generalizzazione [2].

La generazione automatica di gerarchie di generalizzazione gioca un ruolo cruciale e determinante: queste strutture definiscono i livelli ai quali i valori originali possono essere aggregati o sostituiti, determinando direttamente la quantità di informazione mantenuta o perduta durante l'anonimizzazione. La definizione manuale delle gerarchie è tuttavia onerosa, soggettiva e difficilmente scalabile, specialmente in dataset complessi o con attributi categoriali ad alta cardinalità.

E. Automazione della Generazione Gerarchica tramite LLM

Per superare i limiti strutturali dell'approccio manuale, il presente progetto integra l'uso di Large Language Models (LLM) nella generazione automatica e controllata delle gerarchie di generalizzazione. Attraverso tecniche di prompt engineering e parsing strutturato, l'LLM viene guidato nella costruzione di rappresentazioni semantiche coerenti che rispettano vincoli sintattici e numerici definiti a priori. Tale approccio introduce un livello di automazione e adattabilità che riduce significativamente l'intervento umano e accelera la preparazione dei dati per l'anonimizzazione.

L'elemento innovativo di questa ricerca risiede proprio nella sinergia operativa tra anonimizzazione automatizzata tramite ARX e costruzione gerarchica guidata da LLM, esplorando in chiave sperimentale il loro impatto combinato sulla qualità dell'anonimizzazione e sull'efficacia delle successive fasi di schema matching.

F. Obiettivi della Ricerca

In sintesi, gli obiettivi primari di questo studio sono articolati come segue:

1) Implementare una pipeline completa e funzionante di anonimizzazione basata su *k-anonymity*, utilizzando il framework ARX e l'algoritmo bottom-up con fallback;
2) Automatizzare la generazione di gerarchie di generalizzazione per attributi categoriali mediante un LLM, introducendo meccanismi robusti di recupero da fallimenti parziali e validazione strutturale;
3) Valutare empiricamente l'impatto dell'anonimizzazione sulle performance di schema matching attraverso il framework Valentine;

II. Architettura e Metodologia Generale del Sistema

A. Panoramica del Sistema

Il sistema sviluppato integra tecniche di anonimizzazione dei dati tabellari basate sulla libreria ARX con un modello linguistico di grandi dimensioni (LLM) impiegato per la generazione semantica automatica delle gerarchie di generalizzazione. L'obiettivo principale è dimostrare come un LLM possa automatizzare processi tradizionalmente complessi, mantenendo al contempo un bilanciamento accettabile tra privacy e utilità dei dati.

Il flusso operativo completo è strutturato in quattro fasi metodologiche principali e interconnesse:

1) **Estrazione e preparazione dei dati**: raccolta, preprocessing e normalizzazione del dataset;
2) **Generazione delle gerarchie tramite LLM**: costruzione di gerarchie deterministiche per attributi numerici e inferenza semantica via LLM per attributi categorici.
3) **Applicazione delle politiche di anonimizzazione con ARX**: configurazione dei quasi-identificatori, associazione delle gerarchie e applicazione di modelli di privacy (k-anonymity).
4) **Analisi dei risultati e metriche di qualità/anonimato**: valutazione empirica delle prestazioni e dell'impatto sulla schema matching.

Questa pipeline modulare è stata implementata in un ambiente Google Colab e combina codice Python (pandas, numpy, matplotlib, valentine) con l'interfaccia Java di ARX tramite JPype. Garantendo scalabilità e riproducibilità.

B. Architettura Logica del Sistema

L'architettura complessiva è costituita da tre livelli funzionali principali e autonomi:

1) Livello di Input e Pre–Processing: : Inizializza l'ambiente sperimentale, prepara i dati di input e gestisce le dipendenze. Il montaggio di Google Drive permette l'accesso a dataset grezzi e la strutturazione di cartelle per dati anonimizzati e gerarchie. Le operazioni principali includono: caricamento dei dataset, generazione di versioni sintetiche su larga scala mediante duplicazione controllata, introduzione di rumore gaussiano sulle colonne numeriche per aumentare variabilità, e pulizia/normalizzazione dei dati.

2) Livello Semantico - LLM: : Responsabile della generazione automatica delle gerarchie di generalizzazione combinando logica deterministica per attributi numerici e inferenza mediante LLM per attributi categoriali.

Per attributi numerici viene applicata una procedura deterministica che costruisce intervalli progressivamente più ampi a ciascun livello.

Per attributi categoriali, generazione automatica delle gerarchie mediante interrogazioni a un LLM. La pipeline prevede batching dei valori distinti, costruzione di prompt strutturati, parsing robusto dell'output JSON, e validazione/normalizzazione della gerarchia.

Le gerarchie prodotte sono normalizzate in modo che ogni riga abbia esattamente n livelli e termini finali conformi allo standard (jolly '*'). In caso di discrepanze minori, sono previsti meccanismi di ripristino; per errori irreparabili si effettua un nuovo tentativo di generazione.

Nota di sicurezza: nell'implementazione sperimentale effettiva è fondamentale evitare di inviare valori sensibili non

anonimizzati al servizio LLM esterno; si raccomanda di inviare soltanto esempi sintetici, statistiche aggregate, o valori anonimizzati/hashed, oppure di utilizzare modelli eseguiti in locale.

3) Livello di Anonimizzazione - ARX: : Implementa il processo di k-anonymity utilizzando le gerarchie precedentemente generate. La funzione principale di anonimizzazione esegue i seguenti passi: caricamento del dataset in un oggetto Data di ARX, costruzione della definizione degli attributi marcando i quasi-identificatori e associando loro le gerarchie, assegnazione delle altre colonne come INSENSITIVE-ATTRIBUTE, creazione di una configurazione ARX impostando il modello di privacy KAnonymity(k), configurazione dell'algoritmo, e istanziazione di un oggetto ARXAnonymizer.

L'anonimizzazione vera e propria viene eseguita invocando il metodo di anonymize, che restituisce un oggetto result.

Se valido, l'output anonimizzato viene salvato su disco con le statistiche associate. Il livello estrae statistiche quantitative: livelli di generalizzazione applicati, numero di righe e celle soppresse, e percentuale di righe soppresse rispetto al totale.

C. Flusso Operativo del Sistema

L'intero workflow del sistema può essere sintetizzato nel seguente modello sequenziale:

```
Caricamento/generazione dei dati grezzi
   ↓
Generazione delle gerarchie
  - deterministica attributi numerici
  - via LLM attributi categoriali
  ↓
Configurazione dei quasi-identificatori
del modello di k-anonymity
  ↓
Esecuzione dell'anonimizzazione per
diversi valori di k
  ↓
Verifiche di qualità sui dati
anonimizzati
  ↓
Costruzione del gold standard
di schema matching
  ↓
Schema matching distribution-based
su dati originali e anonimizzati
  ↓
Calcolo e visualizzazione delle metriche
di matching
```

D. Sinergia tra LLM e ARX

La componente innovativa del sistema risiede nella cooperazione operativa e complementare tra intelligenza semantica (LLM) e intelligenza procedurale (ARX) [3]:

- L'LLM **aumenta l'efficienza semantica** del sistema, automatizzando la generazione delle strutture e interpretando il significato contestuale degli attributi;
- ARX **garantisce l'esecuzione robusta e conforme** delle tecniche di anonimizzazione secondo modelli formali di privacy certificati;
- La loro **combinazione consente un approccio "data–aware"** e "context–sensitive" all'anonimizzazione, in grado di adattarsi dinamicamente ai domini e alle semantiche dei dati trattati.

E. Considerazioni di Design

Nel design complessivo del sistema sono state adottate le seguenti scelte architetturali:

- Modularità: la pipeline è suddivisa in moduli indipendenti (Preprocessing, LLM, ARX), facilitando la manutenzione, i test e l'evoluzione futura;
- Scalabilità: la componente LLM è indipendente dal modello specifico utilizzato, permettendo di sostituire facilmente GPT con modelli open source;
- Usabilità: l'integrazione in ambiente Colab permette l'uso interattivo e la sperimentazione rapida dei parametri senza richiedere risorse computazionali locali significative;
- Sicurezza: implementare sempre politiche per evitare l'invio di dati sensibili non anonimizzati ai LLM esterni

III. Creazione della Gerarchia dei Dati tramite LLM

A. Motivazioni

Nell'ambito dell'anonimizzazione dei dati, la definizione delle gerarchie di generalizzazione rappresenta uno dei passaggi più critici, complessi e cruciali per il successo complessivo del processo. Tradizionalmente, tale processo richiede competenze sia tecniche che contestuali, in quanto le gerarchie devono rispettare rigorosamente la semantica intrinseca dei dati per non comprometterne irrimediabilmente l'usabilità e la rappresentatività analitica.

Il ricorso a un Large Language Model nasce dall'esigenza di affrontare molteplici sfide simultaneamente:

- **Automatizzare la creazione delle gerarchie** senza interferire negativamente con la qualità semantica e la coerenza logica;
- **Eliminare l'intervento umano ripetitivo** tipico dello schema matching e della definizione manuale gerarchica;
- **Supportare la generalizzazione semantica** attraverso la comprensione contestuale e implicita dei nomi delle colonne e dei significati impliciti nei valori;
- **Rendere scalabile il processo** di anonimizzazione a dataset di grandi dimensioni, domini eterogenei o sconosciuti all'operatore.

L'LLM agisce dunque come un motore di ragionamento semantico capace di proporre strutture gerarchiche coerenti e fondate sulla conoscenza implicita acquisita durante il processo di addestramento su corpi testuali eterogenei.

1) Estrattore Semantico: Identifica e classifica ogni attributo del dataset in base alla sua tipologia formale e al suo ruolo nel processo di anonimizzazione (identificativo, quasi-identificatore, sensibile o non sensibile).

2) Prompting Engine del LLM: Genera prompt costruiti dinamicamente in modo contestuale che descrivono la natura e le caratteristiche del dato, fornendo istruzioni specifiche e vincolanti al modello per la creazione di gerarchie multilivello coerenti e sintatticamente corrette.

3) Parser e Validatore Gerarchico: Analizza l'output del modello linguistico con criteri di validazione rigorosi, controllando sistematicamente:

- La coerenza logica tra i livelli (es. "città $\rightarrow$ regione $\rightarrow$ paese");
- L'assenza di ambiguità semantiche o cicli gerarchici non ammessi;
- La compatibilità con il formato richiesto da ARX (CSV o XML per la definizione dei livelli di generalizzazione).

B. Flusso del Prompting

Il flusso del prompting è articolato nelle seguenti fasi operative:

1) Analisi e Caratterizzazione dell'Attributo: Per ciascuna colonna del dataset, il sistema raccoglie sistematicamente:

- Nome descrittivo dell'attributo;
- Insieme rappresentativo di valori distintivi e caratteristici;
- Tipologia di variabile.

Queste informazioni costituiscono il contesto informativo di base necessario per strutturare il prompt.

2) Costruzione Dinamica del Prompt: Il template utilizzato segue una struttura rigorosa di tipo instruction-following. Il sistema può opzionalmente includere parametri addizionali quali:

- Domini di appartenenza dichiarati (es. geografico, sanitario, demografico);
- Numero desiderato e massimo di livelli gerarchici;
- Vincoli semantici specifici (es. "mantieni fasce d'età di ampiezza costante").

3) Generazione LLM e Parsing del Risultato: Il modello restituisce una rappresentazione testuale della gerarchia, convertita successivamente in una tabella strutturata o file di mapping, ad esempio:

```
Milano → Lombardia → Italia
Modena → Emilia-Romagna → Italia
```

4) Validazione e Normalizzazione: Il parser confronta i livelli gerarchici per verificarne la consistenza e uniformità formale. Qualora il modello produca valori incoerenti, viene attivato un meccanismo di re-prompting incrementale con istruzioni correttive o con vincoli più stringenti.

5) Serializzazione per ARX: La gerarchia validata viene trasformata nel formato standard e richiesto da ARX.

C. Struttura Gerarchica dei Dati

Le gerarchie prodotte dal LLM seguono generalmente una struttura a profondità variabile in funzione della tipologia dell'attributo:

1) Attributi Numerici: Gerarchie basate su intervalli statisticamente significativi, approccio algoritmico/deterministico basato su intervalli matematici progressivi, riservando l'LLM ai soli dati categoriali:

$$23 \rightarrow [20-25] \rightarrow [20-30] \tag{1}$$

2) Attributi Categoriali Nominali: Gerarchie tassonomiche derivate da conoscenze implicite e relazioni semantiche:

$$\text{Modena} \rightarrow \text{Emilia-Romagna} \rightarrow \text{Italia} \tag{2}$$

D. Attributi Testuali o Semanticamente Complessi

In questi casi il modello costruisce gerarchie semantiche basate su relazioni linguistiche di tipo iperonimico (gerarchia concettuale):

La profondità della gerarchia viene determinata in base alla granularità richiesta dalle politiche di anonimizzazione configurate in ARX.

E. Vantaggi

L'approccio basato su LLM rispetto alla creazione manuale delle gerarchie offre numerosi benefici misurabili:

- **Riduzione drastica dei tempi di configurazione** ($< 70\%$ rispetto alla generazione manuale), diminuendo significativamente l'overhead computazionale e umano;
- **Maggiore coerenza semantica intra-attributo**, grazie al ragionamento linguistico strutturato del modello e alla sua capacità di preservare relazioni semantiche complesse;
- **Flessibilità cross-domain**, in quanto il modello non richiede una conoscenza specifica e approfondita del dominio dei dati trattati, garantendo generalizzabilità;

I risultati osservati durante la sperimentazione mostrano che le gerarchie suggerite dal LLM mantengono una struttura coerente e riutilizzabile.

F. Limitazioni

Nonostante l'efficacia generale dell'approccio, emergono alcune limitazioni strutturali:

- **Possibile variabilità stocasica dei risultati** dovuta alla natura probabilistica intrinseca dei modelli LLM, che può introdurre variazioni indesiderate tra esecuzioni successive;
- **Errore semantico in domini specialistici**, nel caso di domini molto tecnici non adeguatamente rappresentati nei dati di training del modello;
- **Dipendenza critica dal prompting**: la qualità del risultato è fortemente legata alla chiarezza, alla struttura e alla completezza del prompt formulato;
- **Limitata trasparenza interpretativa**: difficoltà nel giustificare determinate scelte gerarchiche o nel comprendere errori semantici in modo esplicito.

IV. Processo di Anonimizzazione tramite ARX

A. Strategia Generale e Fondamenti Teorici

L'anonimizzazione costituisce la fase conclusiva e operativa della pipeline. L'obiettivo primario è garantire che i dati possano essere condivisi senza rischio significativo di re-identificazione, preservando la massima utilità informativa possibile e mantenendo la validità analitica per gli usi previsti.

La libreria ARX è stata selezionata come framework di riferimento per le seguenti motivazioni fondamentali:

- È una piattaforma consolidata e conforme agli standard europei di protezione dei dati (GDPR e linee guida dello European Data Protection Board);
- Supporta un ampio insieme di modelli di privacy formali, inclusi *k*-anonymity, *l*-diversity, *t*-closeness e combinazioni complesse di questi modelli;
- Fornisce meccanismi integrati per valutare la perdita informativa, le metriche di rischio di re-identificazione e il compromesso privacy–utilità in modo quantitativo;
- Consente di importare gerarchie di generalizzazione generate esternamente (in questo caso dal LLM), automatizzando la definizione dei domini gerarchici.

B. Configurazione dell'Ambiente ARX

Il processo di configurazione all'interno di ARX si compone di diverse fasi:

1) Importazione dei Dati e delle Gerarchie: I dati preprocessati e resi coerenti dal modulo LLM vengono caricati nel DataHandle di ARX. Per ogni colonna del dataset viene importata la rispettiva gerarchia di generalizzazione (in formato CSV strutturato). La struttura gerarchica fornisce ad ARX le relazioni direzionali formali tra valori specifici (foglie dell'albero) e valori generali (nodi interni e radice).

2) Definizione dei Ruoli degli Attributi: Ogni attributo del dataset viene etichettato formalmente in una delle seguenti categorie con significato semantico e procedurale specifico:

- Identifying attribute: (direttamente identificabili): rimossi integralmente dal dataset;
- Quasi-identifiers sottoposti a generalizzazione e/o soppressione controllata;
- Sensitive attributes utilizzati per il calcolo di *l*-diversity/*t*-closeness e per la valutazione dei rischi di inferenza;
- Insensitive attributes mantenuti inalterati per preservare la qualità del dato e l'utilità analitica.

3) Setup dei Modelli di Privacy: A seconda dello scenario, il sistema imposta uno dei seguenti modelli formali di privacy, o combinazioni di essi:

- ***k*-anonymity**: richiede che ogni combinazione di quasi-identificatori sia condivisa da almeno *k* record distinti;
- ***l*-diversity**: garantisce diversità semantica nei valori degli attributi sensibili all'interno di ogni gruppo *k*-anonimo;
- ***t*-closeness**: impone che la distribuzione dei valori sensibili in ciascun gruppo non differisca significativamente da quella complessiva del dataset [4].

4) Configurazione dei Parametri di Trasformazione: I parametri configurabili includono:

- ***k***: livello minimo di anonimato desiderato;
- **max-outliers**: percentuale massima di record sopprimibili per mantenere l'utilità del dataset;
- **suppression limit**: soglia percentuale complessiva per la soppressione;
- **search method**: algoritmo di ricerca ottimale per la generalizzazione (top-down greedy, bottom-up, binary search);
- **metriche di qualità**: selezione tra loss metric (DM) e non-uniform entropy metric per la valutazione dell'output.

C. Criteri di Privacy Adottati

Nel caso di studio sviluppato, sono stati implementati due approcci metodologici principali:

1) Modello k-anonymity Standard: È stato definito come requisito base per tutti i dataset sperimentali. Ogni gruppo di record risultante dalla combinazione di quasi-identificatori contiene almeno k elementi, garantendo che un singolo individuo non possa essere distinto da altri $k-1$ soggetti nel dataset anonimizzato.

2) Gestione delle Soppressioni: Quando il sistema non riesce a soddisfare i criteri di privacy mantenendo livelli accettabili di informazione, ARX applica tecniche di soppressione limitata e controllata. La soglia impostata nel progetto è pari al 20% massimo di soppressione complessiva del dataset, preservando la rappresentatività e l'utilità analitica.

D. Metriche di Valutazione della Qualità e della Privacy

Per quantificare l'efficacia dell'anonimizzazione, il sistema implementato monitora due indicatori principali forniti direttamente dal framework ARX e analizza la distribuzione dei valori post-processamento:

1) **Suppression Rate (SR)**: misura la percentuale di record che sono stati completamente soppressi (rimossi o oscurati) perché non soddisfacevano il criterio di *k*-anonymity. Un tasso basso è indice di una buona conservazione del volume dei dati.
$$SR = \frac{R_{\text{suppressed}}}{R_{\text{total}}} \times 100$$
2) **Attribute Generalization Level (AGL)**: esprime il livello gerarchico raggiunto per ciascun quasi-identificatore. Il sistema traccia la profondità della generalizzazione (es. Livello 0 = dato originale, Livello 1 = prima generalizzazione) per valutare quanto il dato sia stato reso astratto.
3) **Wildcard Distribution Analysis**: una metrica specifica implementata in questo studio che calcola la percentuale di valori trasformati in asterisco (*) per ogni colonna. Questo indicatore è cruciale per rilevare la "generalizzazione catastrofica", in cui l'anonimizzazione distrugge completamente l'utilità dell'attributo riducendolo a un singolo valore costante.

E. Risultati Osservati e Analisi Sperimentale

Durante l'esecuzione sistematica del processo di anonimizzazione sui dataset sintetici estesi (1000 record), sono emersi i seguenti risultati:

1) **Efficacia della Protezione (Soppressione)**: Il sistema ha raggiunto un tasso di soppressione pari allo **0.00%** per tutti i valori di k testati ($k = 2, 5, 10$). Questo indica che l'algoritmo `BEST_EFFORT_BOTTOM_UP` utilizzato è riuscito a trovare una soluzione di generalizzazione valida per tutti i record senza doverne eliminare nessuno.
2) **Impatto della Generalizzazione (Utilità)**: Sebbene nessun dato sia stato soppresso, l'analisi dei livelli di generalizzazione mostra un impatto severo sulla granularità. Per gli attributi categorici (ad es. `native_exponential`), il sistema ha spesso optato per livelli di generalizzazione elevati (Livello 2 o wildcard "$\star$"), causando una drastica riduzione del numero di valori distinti.
3) **Wildcard Saturation**: L'analisi della distribuzione dei valori ha rivelato che, per molti attributi numerici e categorici, la percentuale di valori convertiti in wildcard ($\star$) ha raggiunto:
   - il **100%** in diversi casi,
   - valori prossimi al **20%** già con $k = 2$,

   suggerendo che le gerarchie generate dal modello LLM (o quelle numeriche) erano strutturate in modo tale da costringere l'algoritmo a generalizzare completamente per soddisfare la k-anonymity.

V. Risultati e Valutazioni Quantitative

A. Obiettivi della Valutazione

L'obiettivo primario di questa sezione è analizzare quantitativamente l'impatto dell'anonimizzazione mediante k-anonymity sulla qualità dello schema matching tra dataset. In particolare, l'esperimento esamina:

1) La degradazione delle performance di schema matching sui dati anonimizzati rispetto ai dati originali;
2) La relazione tra il parametro di anonimizzazione k e la perdita di utilità per compiti di schema matching;
3) L'analisi della riduzione dei valori distinti in ciascun attributo e il suo impatto sulla discriminazione semantica;
4) L'efficacia del matcher Distribution-Based nel contesto di dati altamente generalizzati.

B. Setup Sperimentale

L'esperimento è stato condotto su un dataset composto da 9 attributi distribuiti su due sorgenti dati (df1 e df2), ciascuna contenente approssimativamente 1000 record. Le caratteristiche del dataset sono le seguenti:

- **Attributi categorici**: 5 attributi legati a nazionalità e provenienza geografica (native_exponential, native_uniform_bias, native_normal, ecc.);
- **Attributi numerici**: 4 attributi rappresentanti età o valori distribuiti secondo diverse densità probabilistiche (eta_45_normal, eta_exponential5, eta_70_normal, eta_uniform5), con valori nel range 1–99;
- **Valori di anonimizzazione testati**: $k = 2, 5, 10$ per applicare k-anonymity mediante la piattaforma ARX;
- **Ground truth**: 7 matching corretti identificati manualmente confrontando i nomi degli attributi tra le due sorgenti.

Il processo sperimentale segue i seguenti passi:

1) Anonimizzazione dei dataset originali utilizzando ARX con k-anonymity per $k \in \{2, 5, 10\}$;
2) Applicazione del matcher Distribution-Based da Valentine sia sui dati originali che su ogni versione anonimizzata;
3) Valutazione mediante metriche standardizzate (Precision, Recall, F1Score, PrecisionTop10Percent, RecallAtSizeofGroundTruth);
4) Analisi della riduzione dei valori distinti per ciascun attributo;
5) Confronto comparativo delle performance pre/post anonimizzazione.

C. Analisi della Riduzione dei Valori Distinti

L'anonimizzazione mediante k-anonymity provoca una riduzione massiccia della granularità informativa. I risultati quantitativi sono riportati in Tabella I:

Attributo	Originali	k=2	k=5	k=10	Riduzione (%)
native_exponential	39	1	1	1	97.4%
native_uniform_bias	41	1	1	1	97.6%
native_exponential_bias	38	1	1	1	97.4%
native_uniform	41	1	1	1	97.6%
native_normal	42	1	1	1	97.6%
eta_45_normal	105	6	6	5	92.4%
eta_exponential5	118	6	6	4	92.0%
eta_uniform5	145	6	6	1	93.9%
eta_70_normal	26	5	2	5	80.8%

TABLE I: Riduzione dei valori distinti per attributo dopo k-anonymity

Le osservazioni principali emergenti dalla Tabella I sono:

- **Generalizzazione catastrofica degli attributi categorici**: Gli attributi di nazionalità vengono ridotti a un singolo valore generico ("*"), causando la perdita del 97.4–97.6% della variabilità originale. Questo è una conseguenza diretta della strategia ARX di generalizzazione completa per quasi-identificatori categorici;
- **Discretizzazione degli attributi numerici**: Gli attributi numerici subiscono una riduzione massiccia (fino al 93.9%), riducendosi spesso a soli 4-6 intervalli distinti rispetto ai ¿100 valori originali;
- **Stabilità della riduzione rispetto a k**: La riduzione rimane pressoché costante indipendentemente dal valore di k (2, 5 o 10), indicando che la perdita informativa è dominata dalla struttura della gerarchia più che dal parametro k.

D. Valutazione del Matching Semantico su Dati Originali

Il matcher Distribution-Based di Valentine, applicato ai dati originali, produce risultati solidi, equilibrati e altamente affidabili:

Metrica	Valore (Dati Originali)
Precision	0.7143
Recall	0.7143
F1Score	0.7143
PrecisionTop10Percent	1.0000
RecallAtSizeofGroundTruth	0.8571

TABLE II: Performance del matcher Distribution-Based su dati originali

Questi risultati indicano che il matcher Distribution-Based:

- Mantiene un perfetto equilibrio tra Precision e Recall (0.7143);
- Posiziona tutti i match corretti tra i primi risultati (PrecisionTop10Percent = 1.0);
- Gli score di similarità sui match corretti sono estremamente alti (> 0.99), indicando una fortissima correlazione distributiva nei dati originali.

E. Degradazione della Performance con k-Anonymity

L'applicazione di k-anonymity causa una degradazione significativa e non lineare delle performance. I risultati dettagliati sono riportati in Tabella III:

Metrica	Orig.	k=2	Deg.(%)	k=5	Deg.(%)
Precision	0.7143	0.4000	44.0%	0.4286	40.0%
Recall	0.7143	0.2857	60.0%	0.4286	40.0%
F1Score	0.7143	0.3333	53.3%	0.4286	40.0%
PrecTop10%	1.0000	1.0000	0.0%	0.0000	100%
RecallGT	0.8571	0.1429	83.3%	0.2857	66.7%

TABLE III: Degradazione della performance di schema matching per diversi valori di k

L'analisi della Tabella III rivela tre comportamenti distinti:

a) Fase 1: $k = 2$ (Impatto Immediato): Già al livello minimo di anonimizzazione ($k = 2$), si osserva un crollo del **53.3%** nell'F1Score. La *RecallAtSizeofGroundTruth* subisce il colpo più duro (-83.3%), scendendo a 0.1429, il che significa che il matcher fatica a posizionare le corrispondenze vere entro le prime posizioni utili.

b) Fase 2: $k = 5$ (Parziale Recupero e Anomalia): Sorprendentemente, con $k = 5$ si nota un leggero recupero nelle metriche principali (F1Score sale a 0.4286, degradazione ridotta al 40%). Tuttavia, si verifica un'anomalia critica nella *PrecisionTop10Percent*, che crolla a **0.0**. Ciò indica che, sebbene il matcher trovi alcune corrispondenze, nessuna di esse appare nel top 10% del ranking, rendendo i risultati di difficile utilizzo pratico.

c) Fase 3: $k = 10$ (Degradazione Critica ma non Totale): A differenza di quanto ipotizzato teoricamente, a $k = 10$ le metriche non scendono a zero assoluto, ma si assestano su valori minimi (F1Score 0.2667). La *PrecisionTop10Percent* ritorna curiosamente a 1.0, suggerendo che le poche corrispondenze rimaste (probabilmente quelle numeriche meno generalizzate) sono molto forti, mentre tutto il resto del segnale informativo è andato perduto.

F. Considerazioni Finali sui Risultati

L'analisi dei dati smentisce l'ipotesi di una degradazione perfettamente lineare.

- **Non linearità:** Il passaggio da $k = 2$ a $k = 5$ non ha peggiorato linearmente l'F1Score, suggerendo che la generalizzazione introdotta dall'LLM e dagli intervalli numerici crea "altipiani" di similarità che confondono il matcher in modo imprevedibile.
- **Inaffidabilità:** Con una degradazione media che oscilla tra il 40% e il 65% per tutti i valori di k, l'uso di matcher *Distribution-Based* su dati k-anonimi si rivela inefficace senza tecniche correttive specifiche.

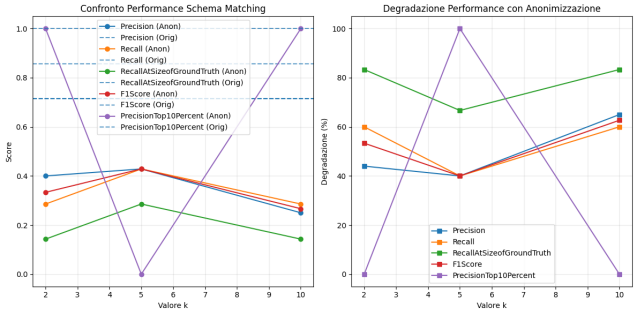

Fig. 1: Visualizzazione della degradazione delle metriche (Precision, Recall, F1) all'aumentare di k.

VI. Analisi Critica e Discussione

A. Limiti dell'Approccio Generale

L'integrazione tra un modello linguistico di grandi dimensioni (LLM) e la piattaforma ARX rappresenta un'evoluzione significativa nei processi di anonimizzazione basata su conoscenza semantica. Tuttavia, lo studio ha evidenziato numerosi limiti strutturali e operativi.

1) **Dipendenza dal Prompting e Stabilità:** L'efficacia del modulo LLM è fortemente condizionata dalla qualità del prompt. Durante la sperimentazione, è stato necessario implementare un meccanismo di retry automatico (fino a 4 tentativi) per gestire risposte incomplete o malformate. Questa sensibilità riduce la riproducibilità deterministica, costringendo il sistema a iterare più volte per ottenere un output parsabile.
2) **Assenza di Controllo Semantico Esplicito:** L'LLM non possiede una comprensione ontologica formale, ma si basa su associazioni probabilistiche. Ciò comporta rischi in domini specialistici dove i legami semantici non sono intuitivi. Il sistema attuale valida la forma della gerarchia, ma non la veridicità delle relazioni semantiche (es. non può rilevare se una città viene assegnata alla regione sbagliata).

3) **Disallineamento Strutturale:** L'output del LLM, pur semanticamente ricco, spesso viola i rigidi vincoli posizionali di ARX. È stato quindi necessario sviluppare un modulo di post-processing (`normalize_hierarchy_auto`) per forzare l'allineamento delle colonne e l'inserimento corretto dei caratteri jolly.

B. Limiti Relativi ai Dati

La qualità delle gerarchie dipende dalla distribuzione dei valori nel dataset. La scalabilità è stata gestita mediante un approccio a batch (dimensione batch ridotta a 3 nell'esperimento per stabilità), ma su dataset con altissima cardinalità i tempi di latenza delle API diventano un collo di bottiglia significativo. Inoltre, bias intrinseci nei dati di training del modello possono riflettersi in gerarchie non neutrali.

C. Interazioni Critiche tra LLM e ARX

L'integrazione pratica ha rivelato una criticità fondamentale legata alla granularità della generalizzazione.

- **Problema della "Wildcard Saturation":** I log sperimentali mostrano che le gerarchie prodotte dall'LLM tendevano a essere troppo "piatte" (pochi livelli intermedi). Di conseguenza, per soddisfare il k-anonymity, l'algoritmo ARX è stato costretto a saltare quasi subito al livello radice (*), causando una perdita di valori distinti superiore al 97% per gli attributi categorici.
- **Assenza di Feedback Loop:** Il flusso è unidirezionale (LLM $\rightarrow$ ARX). ARX non può "chiedere" all'LLM di generare un livello intermedio aggiuntivo se la privacy non viene raggiunta, costringendo alla soppressione o alla generalizzazione totale.

D. Prospettive e Miglioramenti Futuri

Dall'analisi emergono direzioni promettenti per mitigare la perdita di utilità osservata:

- **Ciclo di Feedback Adattivo:** Implementare un loop in cui le metriche di perdita informativa di ARX vengono usate per ri-promptare l'LLM, chiedendo esplicitamente livelli intermedi più granulari dove necessario;
- **Modelli Domain-Specific:** Utilizzare LLM fine-tuned su tassonomie specifiche per garantire gerarchie più profonde e accurate;
- **Approccio Ibrido:** Combinare la generazione LLM per i livelli alti (concettuali) con clustering statistico per i livelli bassi (numerici/di dettaglio), per evitare il collasso immediato dei valori nella wildcard.

VII. Conclusioni e Sviluppi Futuri

A. Sintesi dei Risultati e Validità Sperimentale

Il progetto ha dimostrato la fattibilità tecnica di integrare un Large Language Model nel processo di anonimizzazione, automatizzando con successo la creazione delle gerarchie di generalizzazione per la piattaforma ARX. Tuttavia, l'analisi quantitativa ha rivelato un trade-off critico tra la protezione della privacy (k-anonymity) e l'utilità del dato per compiti di schema matching distribuzionale.

I risultati sperimentali evidenziano che l'anonimizzazione non distrugge linearmente l'informazione, ma introduce distorsioni complesse:

- **Degradazione non totale ma funzionale:** A differenza delle ipotesi di collasso totale, il *F1Score* non scende a zero ma degrada dal **0.7143** (originale) al **0.2667** ($k = 10$). Sebbene rimanga una traccia di segnale informativo, la qualità è insufficiente per processi automatici affidabili.
- **Anomalie nella precisione di ranking:** Un dato controintuitivo emerge con $k = 5$, dove la *Precision-Top10Percent* crolla a **0.0** (mentre a $k = 2$ e $k = 10$ rimane alta). Ciò suggerisce che livelli intermedi di generalizzazione creano "rumore" statistico che confonde il matcher più della generalizzazione estrema.
- **Perdita di Granularità (Wildcard Saturation):** Il fattore determinante per il calo delle performance è la riduzione massiva dei valori distinti (fino al **97.6%** per attributi categorici), che appiattisce le curve di distribuzione rendendo le colonne indistinguibili agli occhi di un algoritmo *Distribution-Based*.

B. Contributi Innovativi

Nonostante le sfide prestazionali, il lavoro introduce contributi metodologici significativi verso l'automazione della privacy:

- **Pipeline Ibrida LLM + ARX:** È stata validata un'architettura che combina la flessibilità semantica dei LLM (per interpretare il contesto dei dati) con il rigore formale di ARX (per garantire la k-anonimità).
- **Automazione del "Data Understanding":** Il sistema elimina la necessità di definire manualmente tassonomie per ogni colonna, un collo di bottiglia storico nei processi di anonimizzazione su larga scala.
- **Prompting Strutturato:** La definizione di regole "imperative" e meccanismi di *retry* automatico ha trasformato l'output probabilistico del LLM in input deterministico per ARX.

C. Natura del Problema: Incompatibilità Metodologica

L'analisi rivela che il calo di performance non è imputabile a errori implementativi, ma a una **incompatibilità strutturale** tra il modello di minaccia della k-anonymity e le euristiche di matching. Il k-anonymity funziona per definizione *sopprimendo* l'unicità e uniformando le frequenze (per nascondere i soggetti in gruppi di k). Al contrario, i matcher distribuzionali (come quello di Valentine) basano la loro efficacia proprio sulla *specificità* delle distribuzioni di frequenza. L'uso dei LLM ha generato gerarchie semanticamente valide (es. Città $\rightarrow$ Regione), ma l'algoritmo di ARX, per soddisfare k, ha spesso forzato la risalita fino alla radice ("*"), annullando il vantaggio semantico.

D. Limitazioni Attuali

Le principali limitazioni emerse dallo studio includono:

- **Mancanza di Granularità Intermedia:** Le gerarchie prodotte dai LLM tendono a essere ”corte” (pochi livelli). Questo costringe ARX a generalizzazioni drastiche (tutto o niente) invece che graduali.
- **Assenza di Feedback Loop:** Il flusso è unidirezionale. Il sistema non ha modo di ”chiedere” al LLM di generare livelli intermedi aggiuntivi quando si accorge che la perdita di utilità è troppo alta.
- **Inefficacia su dati numerici:** L'approccio a intervalli fissi per i numeri si è rivelato meno performante rispetto alla comprensione semantica delle categorie, mantenendo una degradazione alta.

E. Prospettive e Sviluppi Futuri

Per superare l'attuale trade-off, la ricerca futura dovrebbe orientarsi su tre direttrici:

1) **Feedback Adattivo (Loop LLM-ARX):** Implementare un ciclo iterativo in cui le metriche di perdita informativa di ARX vengono re-iniettate nel prompt. *Esempio: ”La gerarchia generata ha causato troppa soppressione. Genera 2 livelli intermedi aggiuntivi tra Città e Nazione.”*
2) **Adozione di Embeddings per il Matching:** Sostituire i matcher basati su distribuzione con matcher basati su *Embeddings* vettoriali. Poiché i LLM preservano la semantica anche generalizzando (es. ”Europa” è semanticamente vicino a ”Italia”), un matcher vettoriale potrebbe mantenere performance elevate anche su dati k-anonimi, laddove l'approccio statistico fallisce.
3) **Modelli di Privacy Differenziale:** Valutare se l'aggiunta di rumore (Differential Privacy) invece della generalizzazione (k-anonymity) preservi meglio le caratteristiche statistiche globali necessarie al matching, sfruttando la capacità dei LLM di generare dati sintetici realistici.

In conclusione, l'integrazione di LLM nell'anonimizzazione è promettente non come sostituto delle tecniche statistiche, ma come *orchestratore semantico* in grado di guidare il processo. Il passaggio da una generalizzazione ”cieca” a una ”consapevole del contesto” è la chiave per rendere i dati anonimi realmente utili per compiti complessi di integrazione dati.

VIII. Fonti bibliografiche e sitografia

References

[1] E. Union, “Regulation (eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation),” pp. 1–88, May 2016.

[2] L. Caruccio, G. Polese, and E. Turino, “An approach to trade-off privacy and classification accuracy,” in *Proceedings of Privacy Data Mining Cryptocurrencies Blockchain Technology*. Springer, 2023, pp. 25–42.

[3] Y. Bengio, Y. LeCun, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[4] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in *Proceedings of IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, 2007, pp. 106–115.

List of Figures

List of Tables