iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XII **Month of publication:** December 2024

DOI: https://doi.org/10.22214/ijraset.2024.65798

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XII Dec 2024- Available at www.ijraset.com

Graph-Based Ranking of Quasi-Identifier Combinations for Privacy Risk Assessment: A PageRank-Driven Framework

Krish Chawla[1], Adarsh Shrivastava[2], Aryan Parmar[3]
Netaji Subhas University of Technology, Dwarka, Delhi, 110078, India

Abstract: ***In the era of data-driven applications, ensuring privacy through effective identification of quasi- identifiers has become a critical challenge. This research focuses on the novel task of rank- ing combinations of quasi-identifiers within a knowledge graph to enhance privacy-preserving mechanisms. Leveraging PageRank as the foundational centrality measure, we introduce modi- fications such as logarithmic scaling to mitigate bias towards general entity nodes that are high in centrality and combination-based averaging to capture multi-entity interactions. The proposed approach not only identifies sensitive quasi-identifier combinations but also addresses limitations of traditional centrality measures such as degree centrality by balancing global significance with local contextual importance. To demonstrate its efficacy, we construct a domain-specific knowl- edge graph from the IMDb dataset, representing entities such as movies, actors, and genres, and validate our method through comparative analyses with existing centrality metrics. Our results highlight the improved accuracy and contextual relevance of the rankings, showcasing the poten- tial for applications in privacy preservation, data anonymization, and knowledge-based analytics. This work bridges the gap between graph-based centrality measures and domain-specific privacy requirements, offering a robust framework for sensitive information management.***

I. INTRODUCTION

In today's digital world, sensitive information, such as personal identities, organizational data, and medical records, is increasingly exposed in various documents. The rise of big data over the Internet has increased the risk of sensitive information getting exposed or accessed by unau- thorized entities and used in fraudulent ways. For example, industries like healthcare, finance, and government agencies handle large volumes of sensitive information that, if compromised, could lead to severe consequences such as data breaches,regulatory penalties, and the erosion of trust. With an increase in the number of cyberattacks that specifically focus on getting access to important data, it has become important to develop new ways to protect important data and keep it safe from malicious attacks. The problem is compounded by the limitations of traditional masking techniques, which of- ten treat all sensitive information equally without considering the context. These methods may fail to preserve the document's readability and utility by either over-masking or under-masking information. Sometimes, masking too much information can compromise the originality of the document and can alert the attackers about the use of masking for cyber deception. For in- stance, in medical documents, masking too many entities can remove critical context, hindering decision-making processes for doctors and researchers. Traditional anonymization techniques such as k-anonymity, l-diversity, and t-closeness have laid the foundation for data privacy by gen- eralizing or suppressing quasi-identifiers to protect sensitive information. While these methods have been instrumental in mitigating re-identification risks, they operate on predefined quasi- identifiers and lack a nuanced understanding of the contextual relationships between data ele- ments. For instance, k-anonymity ensures that a dataset contains groups of k indistinguishable records based on quasi-identifiers but does not account for the interplay between these identifiers or their contextual significance. Similarly, l-diversity and t-closeness add diversity and distribu- tional safeguards, but their reliance on statistical attributes often falls short in scenarios involving complex relationships. A critical challenge in privacy preservation lies in addressing the concept of quasi-identifiers, which are combinations of seemingly non-sensitive attributes that, when linked, can lead to the re-identification of sensitive information. Traditional methods fail to ac- count for the contextual working of these combinations as they focus more on their statistical aspects. For example, in a movie review data set, attributes such as genre or language may not appear important in isolation. However, when viewed in the context of a specific movie, actor or director, their combination could inadvertently personal information or other confidential details. This project aims to bridge this gap by utilizing a risk-based ranking approach that evaluates the contextual importance of entity combinations within a knowledge graph.

©IJRASET: All Rights are Reserved

This project seeks to address these challenges by integrating Knowledge Graphs (KG) and advanced entity-ranking algorithms to efficiently identify quasi-identifiers in a way that maintains context and document integrity. Knowledge graphs represent a graphical view of the entities and relationships between them, capturing both global and local structural dependencies. By integrating graph-based cen- trality measures and a modified version of PageRank, the proposed methodology ensures that sensitive entities and their combinations are identified and anonymized in a way that maintains the readability and utility of the document.

To validate the effectiveness of our approach, we use Spearman's rank correlation coefficient and Kendall's Tau to measure consistency and agreement between our proposed ranking mecha- nism and traditional methods. These metrics allowed us to quantify improvements in capturing the contextual importance of entities while minimizing biases. By comparing the risk scores generated through our bias-reduced method with those of traditional PageRank, we demonstrated a significant improvement in ranking quality and its alignment with domain-specific expectations. Furthermore, the research closely aligns with established privacy principles, such as k-anonymity and l-diversity while adding an extra layer of improvement over them. The knowledge graph layer makes sure that all the domain related information is captured and the combinations of quasi-identifiers are discovered dynamically rather than relying on static-assumptions, offering greater flexibility and adaptability. This is particularly relevant in domains like cyber deception and fake document generation, where the preservation of document readability and originality is paramount.

II. LITERATURE REVIEW

The field of anonymization and privacy-preserving techniques has gained significant atten- tion with the proliferation of data-driven applications.This review explores foundational works, state-of-the-art advancements, and gaps in the literature relevant to this research, focusing on knowledge graph-based approaches, ranking methodologies, and anonymization techniques.

This paper addresses a significant limitation in traditional privacy-preserving data mining ap- proaches: the rigid separation of quasi-identifiers (QIDs) and sensitive attributes. It introduces the concept of sensitive QIDs, which are attributes that simultaneously possess features of both QIDs and sensitive attributes. This nuanced perspective reflects real-world complexities, where attributes like age, job, or address can serve as identifiers and also carry sensitive implications depending on the context. The authors propose innovative privacy models, (l1, ..., lq)-diversity and (t1, ..., tq)-closeness, tailored to handle sensitive QIDs. These models extend traditional metrics like diversity and closeness, allowing for finer-grained control over privacy in datasets. [1]

This paper addresses the critical need for automated anonymization of clinical text to enable the secure sharing of health records while preserving patient privacy. It introduces a novel ap- proach leveraging text embeddings derived from a de-identified dataset to systematically replace words or sentences in clinical notes, ensuring the removal of sensitive information. By evaluating various embedding techniques and models using recently proposed metrics, the study highlights a trade-off between anonymization sensitivity and the retention of relevant medical information. Notably, the results indicate that sentence replacement excels at preserving medical context, whereas word replacement is more effective in achieving higher sensitivity in anonymization. [2]

This paper introduces a novel centrality measure called Isolating-Betweenness Centrality (ISBC), designed to quantify the impact of nodes in complex networks, such as those used for in- formation transmission, epidemic prevention, and infrastructure resilience. Traditional centrality measures like degree centrality, betweenness centrality, and eigenvector centrality typically focus on either local or global aspects of the network structure, such as connectivity, communication paths, and influence propagation dynamics. However, they do not effectively balance time com- plexity and spreading efficiency in capturing both local and global node impacts. The ISBC mea- sure combines two key concepts: Betweenness Centrality (which evaluates the extent to which a node acts as a bridge in the network) and Isolating Centrality (which measures the node's isolation or its ability to affect global influence dynamics). By integrating these two aspects, ISBC provides a more comprehensive assessment of a node's impact, considering both local and global structural influences. The paper tests the performance of ISBC using SIR (Susceptible- Infected-Recovered) and IC (Independent Cascade) epidemic models, comparing it against con- ventional and recent centrality measures on real-world datasets. The results demonstrate that ISBC improves spreading efficiency over both traditional and recent centrality measures while maintaining moderate time complexity, making it a more efficient tool for analyzing node impact in complex networks. [3]

This paper tackles the challenge of privacy-aware sharing of network data, emphasizing the complexities introduced by the interconnected nature of networks. A key contribution is the iden- tification of critical aspects for selecting privacy measures, such as the desired type of privacy, attacker scenarios, data utility, output preferences, and computational feasibility. The authors provide a systematic overview of existing approaches and focus on k-anonymity-based measures that account for the structural surroundings of a node. Through theoretical analysis and empiri- cal experiments on large-scale real-world networks, the study reveals that measures considering a wider node vicinity but leveraging minimal structural information strike an optimal balance between effectiveness and computational efficiency. This finding offers valuable guidance for safely sharing network data while minimizing disclosure risks.

This paper introduces **BackMC**, a simple yet optimal algorithm for local PageRank estimation in undirected graphs. Extensive experiments on real-world and synthetic graphs demonstrate BackMC's superior performance, combining computational efficiency with practical accuracy, making it a robust solution for local PageRank estimation tasks.[5]

This study addresses the increasing need for anonymizing location information in unstruc- tured text, a growing concern as businesses utilize large amounts of unstructured data. Existing anonymization techniques often overlook location data, leaving vulnerabilities for personal lo- cation estimation attacks. To tackle this, the authors propose a novel method for anonymizing location information using a knowledge graph built from actual geographic information systems. The approach is identify articulation points in the knowledge graph as it's anonymization can cutoff links to a lot of important information. Additionally, the study highlights that organiza- tions and place names with a high frequency in unstructured text are more prone to personal identification, emphasizing the importance of safeguarding such information.[6]

This paper addresses the critical issue of privacy in medical record sharing, particularly focus- ing on the tradeoff between data sharing and patient privacy. The authors highlight the risks of re- identification in de-identified data, which can occur even after applying standard de-identification techniques. The paper examines how certain quasi-identifiers—factors that can be inferred from background knowledge—contribute to the likelihood of re-identification. Through a detailed analysis, the authors estimate the probability of re-identification based on these inferable quasi- identifiers, considering both the type and scope of the information available. The findings provide insights into the impact of various quasi-identifiers on re-identification risks, offering valuable guidance for determining the appropriate level of de-identification needed to protect patient pri- vacy without compromising data utility. [7]

This paper addresses the privacy risks associated with the widespread use of pseudonymized user datasets in personalized recommendation systems. It focuses on the challenge of de- anonymizing users by exploiting the uniqueness of their rating patterns across multiple domains, such as movies, books, and music. The proposed method combines probabilistic record linkage techniques with quasi-identifier attacks, utilizing the Fellegi-Sunter model to assess the likeli- hood that two records correspond to the same user based on the similarity of their rating vec- tors. Through experiments on three publicly available rating datasets, the paper demonstrates the method's effectiveness in cross-dataset de-anonymization tasks, achieving high precision and recall (F1-scores ranging from 0.72 to 0.79). The study also explores various factors influenc- ing de-anonymization performance, such as similarity metrics, dataset combinations, and user demographics, highlighting the vulnerabilities of anonymized data in recommendation systems. [8]

This paper explores the challenge of releasing private, person-specific data in a way that en- sures individuals cannot be re-identified while still maintaining the data's utility for research purposes. The authors propose the use of k-anonymity, a formal protection model that ensures privacy by making it impossible to distinguish an individual's data from at least k-1 other indi- viduals in the dataset. The paper discusses the practical deployment of k-anonymity, outlining the necessary policies to ensure its effectiveness. Additionally, the paper examines the risks of re-identification attacks that can still occur in k-anonymous datasets if these policies are not properly followed. The model forms the foundation of several real-world privacy systems like Datafly, m-Argus, and k-Similar, which provide guarantees of privacy protection when properly implemented. This work is critical in addressing the delicate balance between privacy and the usability of data for research, offering a clear framework for data holders to safely share infor- mation while minimizing privacy risks. [9]

This paper addresses the challenge of identifying quasi-identifiers (QIs) in the context of K- anonymity for privacy-preserving data publishing. The authors point out that the effectiveness of K-anonymity heavily depends on the correct identification of QIs, which are attributes that can separate sensitive data from non-sensitive information. Most existing methods for QI iden- tification either overlook this problem or rely on arbitrary choices, which can compromise both the privacy protection and utility of the anonymized data. The paper introduces a new approach for finding QIs based on a relationship matrix that models the potential connections between published data, sensitive attributes, and external knowledge. The proposed method includes a cut-vertex identification algorithm, which helps to find the necessary and minimal set of QIs. This algorithm improves accuracy and avoids the pitfalls of manual partitioning, and it can be ex- tended to handle situations with multiple sensitive attributes. Experimental results demonstrate that the new algorithm offers better partitioning abilities and lower computational complexity, making it highly suitable for big data publishing scenarios. This work contributes to enhancing the reliability and efficiency of privacy-preserving data publishing using the K-anonymity model. [10]

This paper proposes an approach to anomaly detection in dynamic graph streams, where the structure of the graph evolves over time. The method focuses on detecting abrupt changes in the graph, such as sudden spikes in network attacks or unexpected surges in social media followers.

To achieve this, the authors introduce a modified dynamic "PageRank-with-Decay" al- gorithm, which calculates node importance scores by considering the temporal evolution of the graph at each timestep. The decay factor in the PageRank algorithm helps the model adapt to rapid changes and prioritize recent structural shifts. The approach provides a refined mechanism for detecting anomalies, particularly sudden changes in network patterns. The proposed model outperforms state-of-the-art methods in terms of precision and recall, as demonstrated by exper- iments on a real-world dataset, offering faster and more accurate results for detecting dynamic anomalies. [11]

This paper addresses the issue of anonymizing knowledge graphs (KGs) before publishing them, ensuring that sensitive information is protected while still enabling valuable data sharing. Existing approaches for KG anonymization, such as those based on the k-anonymity principle, typically anonymize the entire dataset at the same level, which may not consider individual users' preferences or the sensitivity of their data. To tackle this, the authors introduce the Personalized k-Attribute Degree (p-k-ad) principle, allowing users to specify their own k-anonymity levels. This ensures that each user's data is anonymized to their desired level of privacy, while still pre- venting adversaries from re-identifying individuals with a confidence higher than 1/k. The pa- per also presents the Personalized Cluster-Based Knowledge Graph Anonymization Algorithm (PCKGA), which implements the p-k-ad principle. This algorithm is designed to anonymize KGs in a way that respects user-specified privacy levels. Experimental results on four real-world datasets show that the PCKGA outperforms previous anonymization techniques, providing bet- ter quality anonymized KGs while maintaining personalized privacy protection. This approach offers a more flexible and user-centric solution to KG anonymization.[12]

This paper addresses a critical issue in the realm of text-based medical applications: the need to balance data anonymization with the preservation of contextual information essential for machine learning model performance. It systematically evaluates the impact of various anonymization techniques on state-of-the-art machine learning models across multiple NLP tasks, providing a comprehensive analysis that is both timely and relevant.[13]

This paper presents KG-Rank, an innovative framework that integrates large language models (LLMs) with medical knowledge graphs (KGs) to address the critical issue of factual consistency in long-form question answering (QA) within the medical domain. The authors effectively tackle a major challenge in applying LLMs in clinical settings by combining KG-based entity retrieval with advanced ranking and re-ranking techniques. The inclusion of ranking mechanisms tailored to the medical domain represents a novel approach, and the paper successfully demonstrates the model's utility through impressive performance improvements—over 18 percent in ROUGE-L score—across multiple medical QA datasets. Extending the framework to open domains like law and history and achieving a 14 percent improvement in ROUGE-L score highlights the adapt- ability and broader applicability of KG-Rank.[14]

This paper addresses the critical issue of balancing privacy and utility in text anonymiza- tion, particularly in the context of mitigating re-identification attacks by large language models (LLMs). LLMs, with their sophisticated ability to memorize and interlink granular details, pose unique challenges to traditional anonymization techniques. A standout feature of the proposed framework is the use of Direct Preference Optimization (DPO) to distill anonymization capabil- ities into a lightweight model suitable for large-scale and real-time scenarios. This is a practical solution addressing the computational challenges of implementing LLM-based anonymization frameworks in resource-intensive environments.[15]

III. PROBLEM STATEMENT

The ever-expanding volume of digital data has made privacy preservation a critical concern across industries such as healthcare, finance, media, and government. Sensitive information, of- ten embedded within datasets, is increasingly vulnerable to exploitation, leading to severe conse- quences such as identity theft, financial fraud, and regulatory violations. Traditional anonymiza- tion techniques, while foundational, fall short in addressing the domain-specific complexities of modern data and the nuanced threats posed by quasi-identifiers (QIs)—combinations of attributes that can be exploited for re-identification.

The identification and effective masking of quasi-identifiers are paramount for:

1) *Preventing Data Breaches:* Unauthorized access to sensitive information through quasi- identifiers poses a growing risk, especially in domains handling large datasets like health- care and entertainment.
2) *Ensuring Compliance:* Stringent regulations like GDPR and HIPAA demand robust meth- ods for preserving privacy while retaining data utility.
3) *Balancing Privacy and Usability:* Overgeneralized anonymization methods often sacrifice data utility, undermining critical applications such as medical research, fraud detection, and public policy planning that may need to retain the utility of the document alongside hiding sensitive information that may lead to malicious practices.

A. *Shortcomings of Existing Systems*

1) Limitations of Traditional Privacy Models: Existing privacy models such as *k*-anonymity, *l*-diversity, and *t*-closeness fail to address relational or contextual links among attributes. Their approach to identifying important entities is static, with predefined entities and their importance. For example, *k*-anonymity ensures that each record is indistinguishable from *k* others based on selected attributes but fails to address relational or contextual links among attributes, leading to overgeneralization, reduction in data granularity, and limited usability in fields requiring detailed analysis.
2) Ineffectiveness of Entity-Centric Approaches: Current systems often rely on individual attributes for masking while ignoring the relational complexity within datasets. Entities are considered as individual units, ignoring the fact that a combination of attributes can dis- close sensitive information. For example, in a movie dataset, masking individual attributes like "actor" or "movie" might suffice, but combinations (e.g., actor + movie + year) could uniquely identify an individual or record
3) Bias in Graph-Based Ranking Techniques: Centrality measures, such as degree or PageRank, are widely used in knowledge graph frameworks for ranking sensitive enti- ties. These methods overemphasize general entities (e.g., popular actors, common genres), overshadowing less frequent but contextually significant entities. This skews rankings and undermines the contextual importance of more specific entities. This inability to balance global significance with local relevance compromises the accuracy of entity ranking and masking strategies.
4) Over-Masking and Under-Masking Dilemmas: Over-anonymization erodes the utility and readability of datasets, making them impractical for real-world applications. For ex- ample, masking too many nodes in medical documents can obscure critical information, hindering decision-making for researchers and practitioners. On the other hand, insufficient protection of sensitive entities leaves datasets vulnerable to adversarial attacks, increasing the risk of re-identification and undermining anonymization effectiveness.
5) Lack of Context-Aware Mechanisms: Existing systems lack dynamic approaches to adapt to the specific context of datasets and treat all attributes with equal sensitivity, ignoring their contextual importance and the risk posed by their combinations. These systems fail to incorporate real-world insights into masking strategies, resulting in a one-size-fits-all approach that is both inefficient and ineffective.

B. *The Need for a New Approach*

The shortcomings of existing methods highlight the urgent need for innovative solutions that:

1) Dynamically identify and rank combinations of different quasi-identifiers based on their re-identification risk within a given dataset.
2) Address biases in entity ranking by balancing the global and local importance of attributes.
3) Integrate contextual awareness through tools like knowledge graphs to provide targeted anonymization that preserves data usability while ensuring robust privacy.

This research seeks to tackle these gaps by focusing on a scalable and contextually aware framework to identify and mask quasi-identifiers effectively, ensuring privacy without compro- mising utility. The solution emphasizes fairness in ranking algorithms and evaluates the impact using robust statistical metrics such as Spearman's Rank Correlation and Kendall's Tau, ensuring transparency and reliability.

IV. METHODOLOGY

A. *Dataset Used*

1) *Dataset Overview:* For the purposes of this study, we utilized the IMDb Movie Dataset, a publicly available collection of comprehensive metadata about movies. This dataset is widely used in research across various domains, such as recommendation systems, natural language processing (NLP), and network analysis. The IMDb dataset contains a wide range of movies and offers a diverse set of attributes such as movie title, actors, directors, genre, language, and much more. This allows for extensive research on the importance of different types of entities in a huge and diverse dataset of over 80,000 movies.
2) *Data Preprocessing:* Prior to utilizing the dataset for knowledge graph construction and subsequent ranking analysis, several preprocessing steps were conducted to ensure the dataset's integrity and relevance. Records with missing or incomplete critical fields were removed to maintain the consistency of the dataset, along with the removal of duplicates to ensure that each movie was represented only once in the dataset.

3) Subset Selection: Although the full IMDb dataset includes over 80,000 movies, this re- search utilized a subset of approximately 10,000 movies. This subset was selected based on the completeness of the metadata, with a focus on movies that had sufficient information for meaningful graph construction and entity extraction. The size of the subset also en- sured computational feasibility while maintaining the diversity of entities and relationships needed for ranking experiments.

4) Application in Knowledge Graph Construction: The IMDb dataset served as the founda- tion for constructing a knowledge graph, in which nodes represent entities such as movies, actors, directors, and genres, while edges represent relationships between these entities (e.g., actors acted in movies, movies belong to genres). This graph structure facilitated the application of graph-based ranking algorithms such as degree centrality or PageRank and the proposed modified ranking method. The IMDb dataset's rich metadata and inherent re- lationships provided a suitable environment for evaluating the effectiveness of these ranking techniques in the context of movie-related entities.

The IMDb dataset served as a cornerstone for this research, providing a rich and diverse reposi- tory of structured information essential for achieving the study's objectives. Its extensive meta- data and inherent relationships allowed for a comprehensive exploration of entities such as movies, actors, genres, and directors, which were critical in constructing a robust and meaningful knowledge graph. The careful preprocessing of the dataset ensured data consistency, relevance, and quality by addressing missing values and removing duplicates, while the subset selection enabled computational efficiency without compromising the diversity and integrity of the data. This balance between dataset richness and practical feasibility proved vital in enabling advanced entity ranking experiments and masking methodologies. Furthermore, the dataset's versatility and adaptability to graph-based analysis have underscored its suitability for exploring innovative techniques in text anonymization and context-aware entity masking. By leveraging its struc- tured nature and inherent relationships, this research has demonstrated the potential for scalable and impactful solutions in privacy-preserving applications. As a widely recognized and reli- able source of information, the IMDb dataset not only supported the goals of this study but also exemplified the value of leveraging real-world datasets in addressing complex challenges at the intersection of natural language processing, network analysis, and data privacy. This section con- cludes with the affirmation that the dataset's characteristics aligned seamlessly with the research needs, facilitating meaningful insights and laying the groundwork for future advancements in this domain.

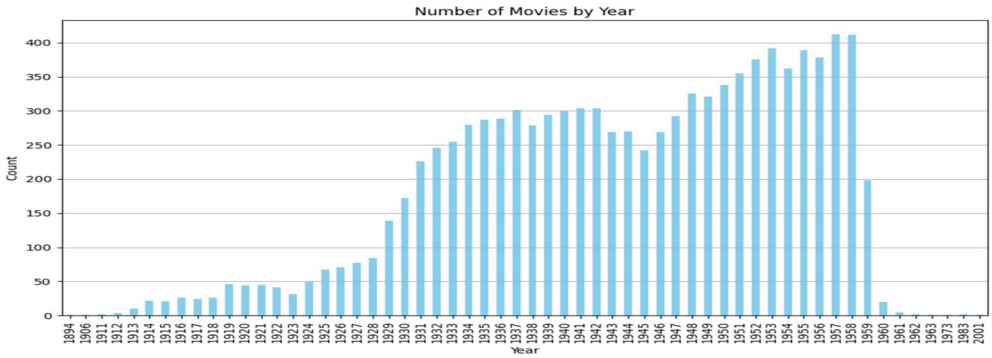


Figure 1: Distribution of metrics for 4001-6000 entries

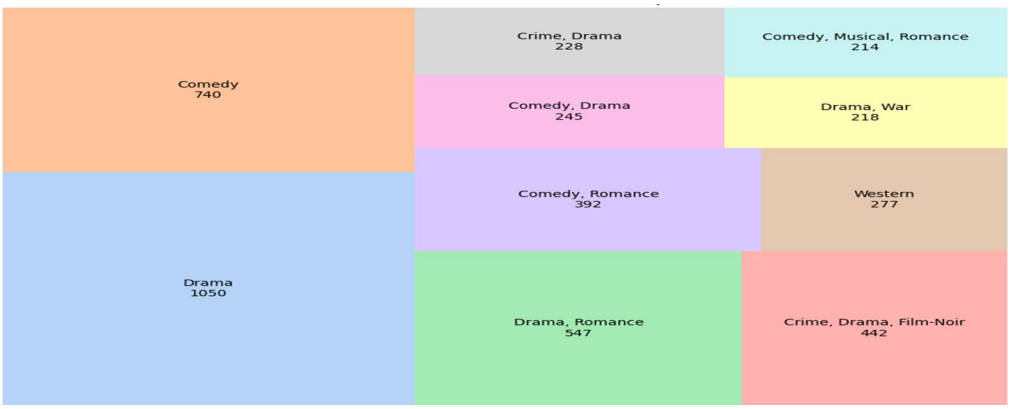


Figure 2: Distribution of metrics for 4001-6000 entries

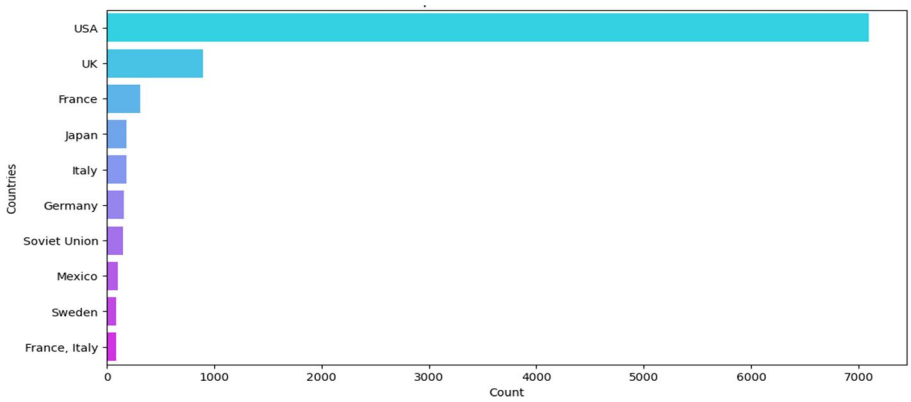


Figure 3: Distribution of metrics for 4001-6000 entries

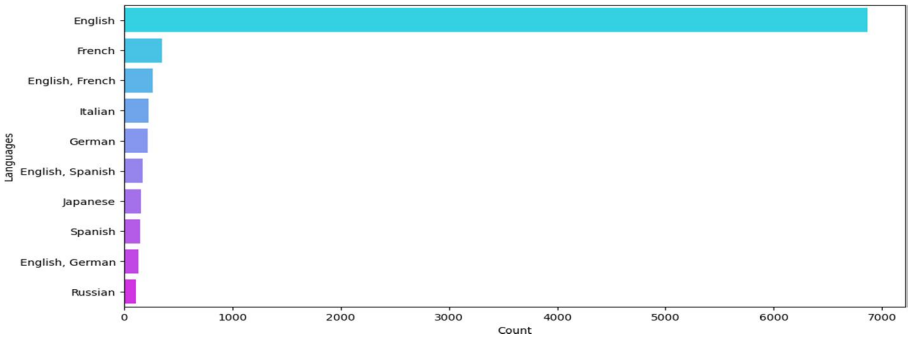


Figure 4: Distribution of metrics for 4001-6000 entries

B. Knowledge Graph Construction

1) *Entity Extraction and Relationship Identification:* Entities are the core building blocks of the knowledge graph and identifying them from the dataset was a crucial task. We focused on several primary entities within the movie domain such as movie names, actors, genre, language, country, and production company. On the other hand, the relationships between entities form the structure of the knowledge graph. These relationships help connect entities and provide context for the analysis and ranking of entities.
2) *Graph Construction Using Neo4j:* We utilized Neo4j AuraDB, a graph database, to con- struct and manage the knowledge graph. Neo4j provides a highly efficient storage and query model for graphs, making it an ideal choice for building large-scale knowledge graphs that need to be queried for entity ranking and centrality calculations. Using Neo4j allows the graph to be scalable and flexible, supporting efficient management of 48,090 nodes and 395,972 relationships present in our knowledge graph.

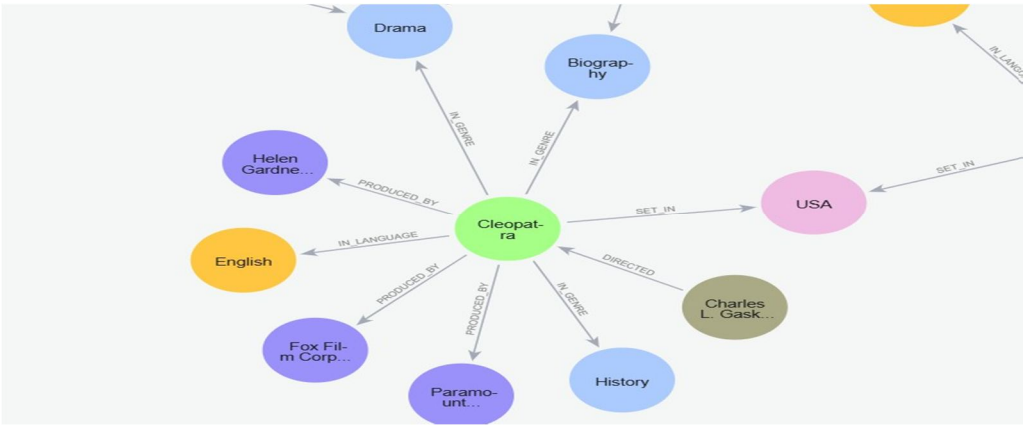


Figure 5: A subgraph of our knowledge graph

Source Node	Relationship	Target Node
Movie	IN˙GENRE	Genre
Director	DIRECTED	Movie
Writer	WROTE	Movie
Actor	ACTED-IN	Movie
Movie	PRODUCED-BY	Production Company
Movie	SET-IN	Country
Movie	IN-LANGUAGE	Language
Actor	WORKED-WITH	Actor
Actor	ASSOCIATED-WITH	Production Company

Table 1: Table of Relationships in the Knowledge Graph

C. Ranking and Risk Scoring

Following the construction of the knowledge graph and dataset preparation, the Ranking and Risk Scoring Methodology emerges as the central analytical framework of this study. Entities and their interrelations, as defined in the knowledge graph, represent structured data where certain entities or combinations can act as quasi-identifiers. These are attributes or sets of attributes that, while not uniquely identifying on their own, can be combined to reveal sensitive information. The Ranking and Risk Scoring Methodology addresses this challenge by assigning a quantified sensitivity score to entity combinations and ranking them accordingly. This method aims to assign risk scores to different combinations of entities systematically to assess and represent their potential to compromise anonymity or reveal sensitive information. This ranking aids in the prioritization of entities for anonymization while ensuring minimal distortion of the dataset's utility. In our ranking system, we choose PageRank as the primary centrality measure for assigning risk scores to entities and their combinations. PageRank is particularly well-suited for this task because it goes beyond local connectivity, capturing the global influence of nodes within the graph. Unlike measures like degree centrality or closeness, which only considers the number of direct connections, PageRank factors in the importance of the nodes that an entity is connected to, making it a more holistic measure of significance. It also holds computational advantage over centrality measures like betweenness centrality or eigen vector centrality and is much better suited for large-scale networks that are sparse in nature. Also, the importance of an entity in re- vealing sensitive information often lies in its position within the global graph structure. PageR- ank's ability to prioritize globally significant nodes aligns well with the objectives of privacy protection and makes it a better and practical choice for the assigning of risk scores. Despite its advantages, raw PageRank scores are often distributed across a very narrow range because of the sparse nature of large-scale networks. This makes it difficult to differentiate enti- ties with slightly varying levels of influence, especially in cases where we have many entities and their combinations. Another drawback is that highly significant nodes often dominate rankings, overshadowing moderately significant combinations that may also represent privacy risks. To address these issues, we enhance the use of PageRank scores through normalization and loga- rithmic scaling, introducing two key improvements:

1) *Normalization:* All PageRank scores are scaled by dividing each score by the maximum PageRank value in the graph. This adjustment ensures that scores are distributed consis- tently across the range [0, 1], making comparisons between entities and combinations more interpretable.
2) *Logarithmic Scaling:* The raw PageRank score for each entity is transformed using a log- arithmic function. This reduces the disproportionate influence of highly ranked entities while amplifying the distinctions between entities with smaller scores. This creates a more balanced representation of risk levels.

The original formula for the PageRank of a node $P(i)$ is as follows:

$$P(i) = \frac{1-d}{N} + d \cdot \sum_{j \in \text{In}(i)} \frac{P(j)}{\text{OutDegree}(j)}$$

Where:

- $P(i)$: PageRank score of node i
- d: Damping factor (usually set to 0.85)
- N: Total number of nodes in the graph
- In(i): Set of nodes linking to i
- OutDegree(j): Number of outgoing edges from node j

To normalize, the raw PageRank scores $P(i)$ are scaled relative to the maximum PageRank value P_{max}:
Where:

$$P_{\text{norm}}(i) = \frac{P(i)}{P_{\text{max}}}$$

$$P_{\text{max}} = \max(P(j)) \quad \text{for all } j \text{ in the graph.}$$

To apply logarithmic scaling:

$$P_{\log}(i) = \log(1 + P(i))$$

This reduces the dominance of high scores and enhances differentiation among lower scores.
For a combination of entities $C = \{e_1, e_2, \ldots, e_k\}$, the combined risk score is calculated as the average of normalized and log-scaled scores:

$$\text{Risk Score}(C) = \frac{1}{2}\left(\frac{\sum_{e \in C} P_{\text{norm}}(e)}{|C|} + \frac{\sum_{e \in C} P_{\log}(e)}{|C|}\right)$$

Where:

- $|C|$: Number of entities in the combination C.

This formula ensures a balanced measure of risk that accounts for both raw influence and adjusted significance, producing more interpretable and meaningful scores for entity ranking in the context of anonymization. By refining the way PageRank scores are utilized, we achieve a risk-scoring framework that effectively balances sensitivity and accuracy. These enhancements are critical for prioritizing quasi-identifiers in text anonymization, ensuring that both high-risk and contextually relevant entities are appropriately addressed. This methodological improvement highlights the power of integrating global centrality measures with practical scaling techniques to advance privacy-preserving data analysis.

D. Evaluation Strategy

The evaluation methodology for our research rigorously assesses the performance and effec- tiveness of our proposed approach to text anonymization using knowledge graphs. To ensure a comprehensive evaluation, we employ both quantitative measures and comparative analysis to validate the performance of our ranking system and risk-scoring method. Our methodology in- volves testing the rankings derived from the entities in the IMDb dataset, which served as the basis for constructing the knowledge graph. Since our approach utilizes a modified PageRank algorithm for assigning risk scores and ranking entity combinations, we establish ground truth rankings using centrality measures such as degree centrality and closeness centrality. These cen- trality measures serve as benchmarks because they quantify the importance of nodes within the graph, which aligns with our premise that central nodes represent more significant entities.To evaluate the validity of our rankings, we calculate Spearman's Rank Correlation Coefficient and Kendall's Tau between the ground truth rankings and the rankings produced by our method. These metrics provide robust insights into the concordance between our proposed rankings and the expected rankings, ensuring that the results are consistent with the theoretical foundation of centrality-based importance in graph networks. In addition to ranking validity, we compare the raw PageRank scores with the modified PageRank scores used in our approach. This analy- sis highlights the improvements introduced by normalization, log scaling, and averaging in the risk-scoring process. Specifically, we focus on how these modifications enhance the distribution of scores, improve differentiation among combinations, and increase interpretability for practi- cal applications. By examining both ranking quality and risk score distribution, our evaluation methodology not only validates the effectiveness of the proposed method but also demonstrates its advantages over traditional approaches. This dual focus ensures that our approach is both theoretically sound and practically useful in the context of text anonymization using knowledge graphs.

V. IMPLEMENTATION DETAILS

The implementation of the proposed ranking and anonymization methodology involved a sys- tematic integration of advanced tools, statistical techniques, and computational strategies. The implementation relied on a robust suite of Python-based libraries to streamline data processing, graph analysis, and statistical validation. Neo4j AuraDB is used as the graph database to store our knowledge graph of 48090 nodes and 395972 relationships as it allows efficient querying of large-scale graphs.

On the other hand, Networkx is used as the library for graph-based com- putations, such as calculating centrality measures and analysing the structure of our knowledge graph along with using the SciPy library to perform statistical analysis of the obtained rank- ings and validating their alignment between the ground truth rankings and proposed rankings. The dataset, containing extensive metadata about movies, was processed iteratively to manage computational costs effectively. The knowledge graph was constructed using a batch processing approach where a dataset of 10,000 entries was divided into 10 batches of 1000 entries to handle the computational costs of running cipher queries on a large dataset.

The ranking process was based on two key measures:

1) *Raw Risk Scores:* These were calculated using centrality metrics like degree centrality, which reflects the number of direct connections a node has within the graph.
2) *Combined Risk Scores:* A refined metric incorporating logarithmic transformations of cen- trality scores to address biases associated with highly connected nodes.

The rationale for using degree centrality as the ground truth for ranking was its intuitive rep- resentation of importance as nodes with high degree centrality are inherently well-connected, making them more likely to be quasi-identifiers or sensitive entities in the context of anonymiza- tion. Also, degree centrality offers a straightforward and interpretable measure, providing a reliable baseline for validating more complex ranking methods. In addition to statistical met- rics, a detailed analysis of specific rows is also conducted to observe changes in raw scores and combined scores were observed for the combinations of entities. The distribution of risk scores was analysed to demonstrate the improvement in risk scores leading to better interpretability and distinction between the score of each entity. The analyses of specific rows allowed to assess the quality of assigned risk scores as well as qualitatively assess the rankings and ensure that important and sensitive combinations of entities are being ranked in a systematic manner. The refinement achieved through combined scores demonstrated improved prioritization, ensuring that rankings were both contextually relevant and aligned with the anonymization objectives.

VI. RESULTS AND DISCUSSION

The results of this study highlight the effectiveness and robustness of the proposed ranking methodology, particularly in the context of anonymization via knowledge graphs. Various sta- tistical metrics, ranking evaluations, and analytical visualizations provide a comprehensive un- derstanding of the system's performance. The statistical evaluation has been done in batches of 2000 entries analysing total of 6000 entries from the dataset. The evaluation is based on the comparison of the ground truth rankings based on degree centrality and the rankings obtained through our proposed method.

Metric	Mean Value	Median Value	Standard Deviation	Minimum Value	Maximum Value
Kendall's Tau	0.958803	0.957013	0.018606	0.842537	0.998816
Spearman's rank correlation coefficient	0.993708	0.994046	0.004838	0.940573	0.999966

Table 2: Statistical metrics for 1-2000 entries

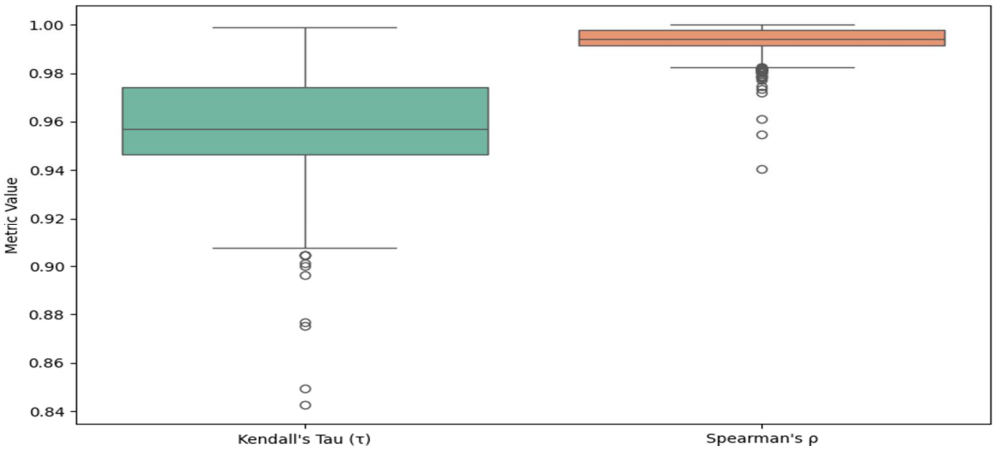


Figure 6: Distribution of metrics for 1-2000 metrics

Metric	Mean Value	Median Value	Standard Deviation	Minimum Value	Maximum Value
Kendall's Tau	0.967706	0.968282	0.013056	0.890088	0.998090
Spearman's rank correlation coefficient	0.995696	0.996143	0.002981	0.964525	0.999924

Table 3: Statistical metrics for 2001-4000 entries

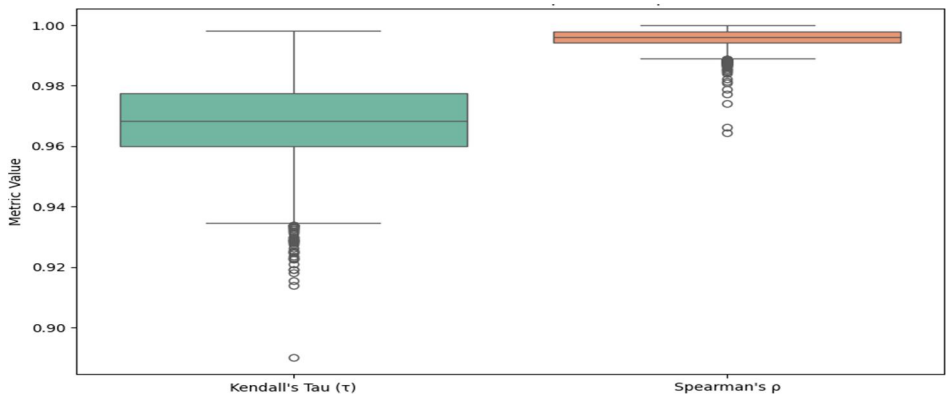


Figure 7: Distribution of metrics for 2001-4000 entries

Metric	Mean Value	Median Value	Standard Deviation	Minimum Value	Maximum Value
Kendall's Tau	0.966275	0.968139	0.015068	0.891197	0.996376
Spearman's rank correlation coefficient	0.995366	0.996085	0.003306	0.973118	0.999939

Table 4: Statistical metrics for 4001-6000 entries

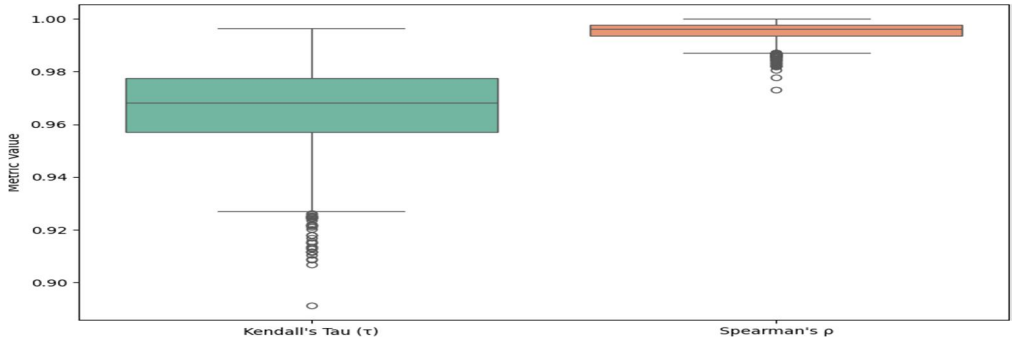


Figure 8: Distribution of metrics for 4001-6000 entries

These values indicate a strong ordinal association between the pairwise comparisons of the proposed rankings and the ground truth. The high mean and maximum values suggest that most rankings are concordant, while the small standard deviation points to consistency across diverse samples. Spearman's rank correlation coefficient confirms a high degree of monotonic correla- tion, demonstrating that the overall ranking order aligns closely with the expected ground truth. This high correlation further reinforces the reliability of the ranking system. Since, the values of the correlation metrics are consistent throughout different batches of entries, it proves the con- sistency of the ranking system and its ability to handle outliers and provide a concrete basis for ranking the combinations. Apart from this, analysis of score distribution and rankings were done on rows individually to observe the change in scores as we use the original algorithm compared to our method. Analysing some of the rows individually also allows us to qualitatively determine the quality of rankings and make sure that right set of entities are given high priority based on their risk scores.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

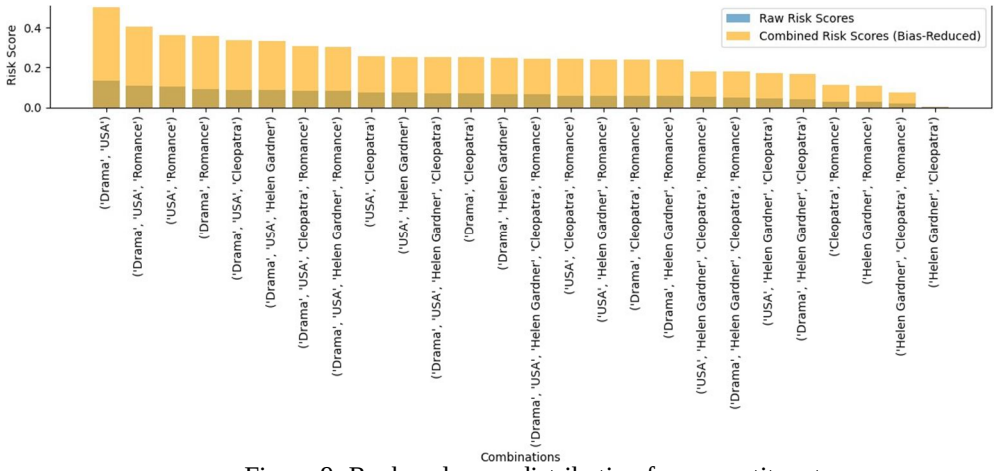


Figure 9: Rank and score distribution for one entity set

Different types of analysis on specific cases highlighted the effectiveness and distinctiveness of our proposed method, particularly in improving the distribution of scores and refining the ranking of entities. Our method demonstrated positive changes in the ranking process by prior- itizing entity combinations that hold significant influence within the knowledge graph. This is largely because PageRank considers not only the centrality of a node but also the strength and significance of its connections with neighboring nodes. For instance, the rank of the entity com- bination (’Drama’, ’Romance’) shifted from 6th place to 4th place under our method. This shift underscores the influence of such combinations in reducing the anonymity of sensitive entities. In a real-world context, the presence of these two genres together can significantly narrow the search space for an attacker, increasing the likelihood of identifying sensitive information related to a specific movie or individual.

By assigning a higher rank to such combinations, our method emphasizes the critical need to anonymize at least one entity in the pair to sever the connection between them. This disconnection disrupts potential inference chains, effectively preventing the re-identification of sensitive entities. This adjustment in ranking reflects our method’s ability to capture the contextual influence of entity combinations within a graph, prioritizing those that pose greater risks to privacy. Overall, this analysis not only demonstrates the practicality and relevance of our approach but also validates its potential to address privacy concerns by identifying and mitigating high-risk connections in real-world scenarios.

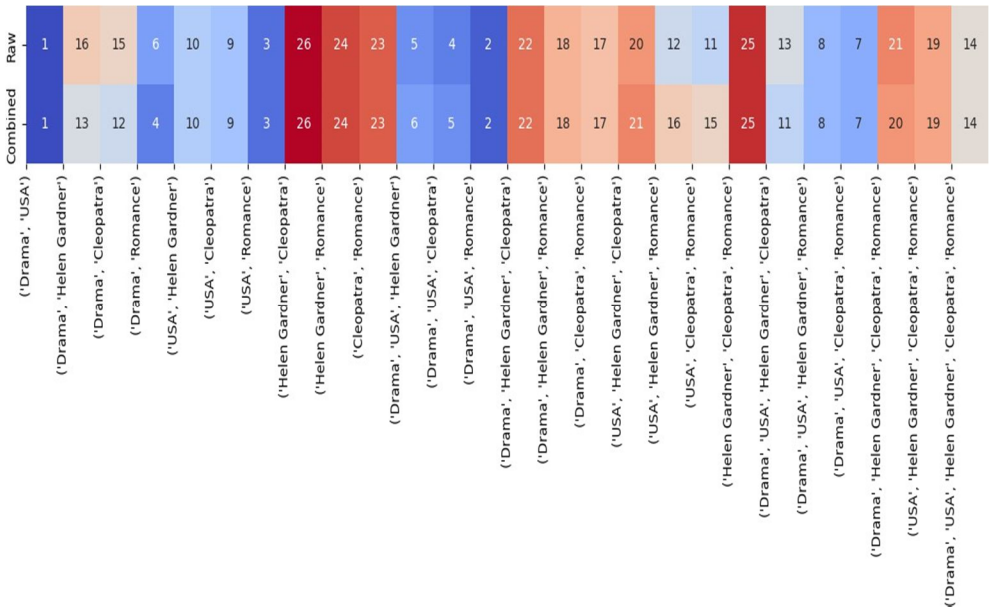


Figure 10: Changes in ranking for one entity set

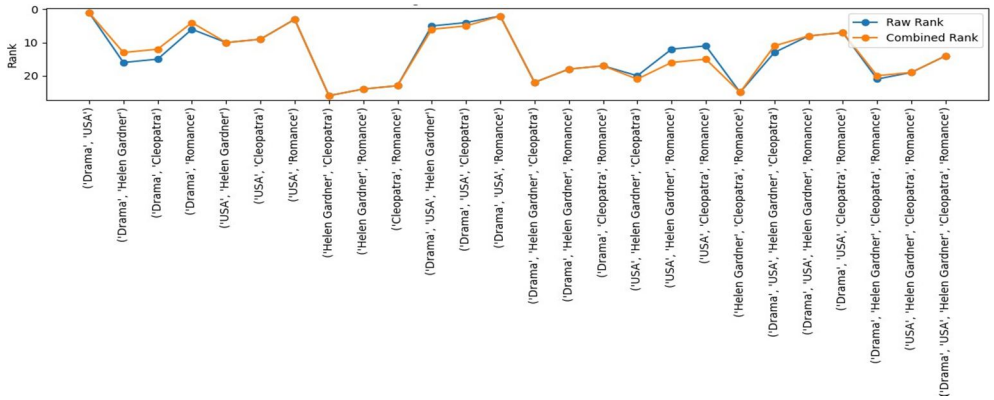


Figure 11: Rank Progression from Raw to Combined Scores

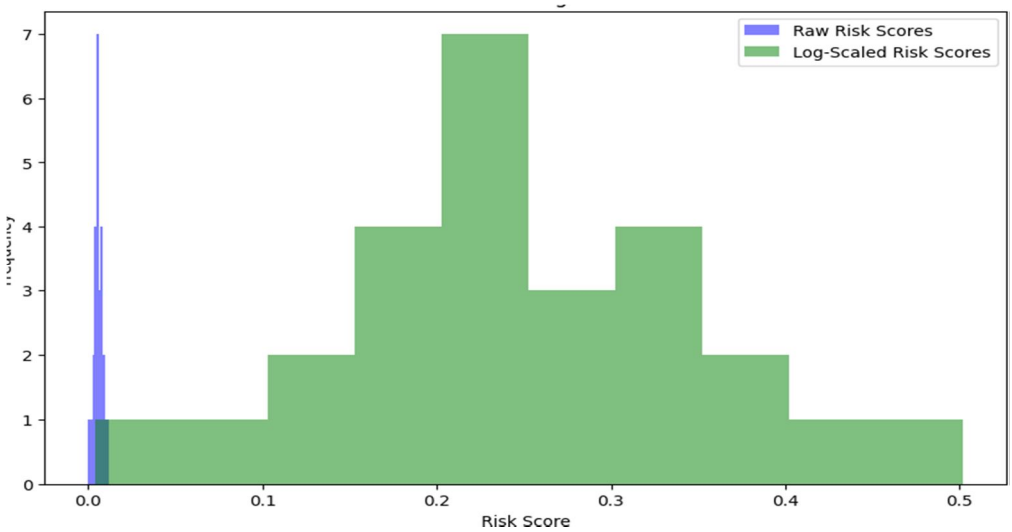


Figure 12: Score Distribution of Raw and Combined scores for one entity set

The results demonstrate the efficacy of the proposed method in addressing critical privacy con- cerns by leveraging the structural properties of the knowledge graph to analyse relationships and centralities dynamically, ensuring that highly sensitive combinations are prioritized for masking. Traditional privacy-preserving techniques such as k-anonymity focus on generalizing or sup- pressing quasi-identifiers to prevent re-identification, often at the cost of significant data utility loss. The comparison of raw risk scores and combined scores reveals a significant shift in rank- ings, emphasizing the importance of contextual connectivity in determining entity sensitivity. Contextually sensitive combinations experienced a positive rank shift due to their higher collec- tive centrality in the graph, underlining the importance of their anonymization to disrupt potential re-identification pathways. This adaptability to context and relational importance differentiates the proposed system from static methods, as it accounts for both the intrinsic significance of nodes and their interconnections. The observed consistency in high Spearman's rank correla- tion coefficent and Kendall's Tau values across all analysed subsets validates the robustness and reliability of the ranking mechanism.

Moreover, the ability to handle large datasets in batches while maintaining computational ef- ficiency further highlights the scalability of the approach. By incorporating the local and global graph metrics, the proposed system ensures a better balance between privacy and utility, pre- serving contextual coherence while anonymizing critical entities. This innovative integration not only enhances data privacy but also broadens the scope of applications, from healthcare and cy- bersecurity to text anonymization in sensitive domains. The method's focus on dynamic masking ensures that the masked output aligns with the sensitivity of the data, maintaining readability and usability. In summary, the proposed system not only addresses the drawbacks of existing meth- ods but also introduces a novel perspective to privacy protection by adapting to the underlying data relationships and their contextual importance. This approach opens new avenues for advancing privacy-preserving systems while retaining data utility, ensuring its applicability across diverse domains.

VII. CONCLUSION

This study introduces a graph-based framework for context-aware anonymization that com- bines PageRank and degree centrality to rank entity combinations dynamically. By leveraging the structure and interconnectedness of entities within a knowledge graph, our method ensures that high-risk combinations, which are most vulnerable to attacks, are prioritized for anonymiza- tion. This approach addresses significant limitations of traditional privacy-preserving systems such as k-anonymity, which often fail to account for contextual relationships and the varying sensitivities of entity combinations. The inclusion of graph-based metrics not only provides a nuanced understanding of the data but also enables the masking of influential entities, ensuring robust privacy protection while preserving the document's readability and contextual integrity.

Our results demonstrate that the proposed method excels at identifying and prioritizing sensi- tive combinations, even in large-scale datasets. By analysing subsets of the dataset in batches, the approach proved computationally efficient and scalable, highlighting its practical applicabil- ity. The method's ability to re-rank combinations based on their contextual importance further underscores its effectiveness in dynamic and complex data environments. In addition to its tech- nical contributions, the study opens new possibilities for applying graph-based privacy systems to various domains, such as healthcare, legal documentation, and cybersecurity. The scalability, adaptability, and robust performance of the approach establish a foundation for developing future anonymization techniques that balance privacy and utility effectively. This work not only high- lights the importance of integrating contextual sensitivity into privacy preservation but also sets a precedent for leveraging graph analytics in creating smarter, more adaptive privacy systems.

VIII. FUTURE WORK

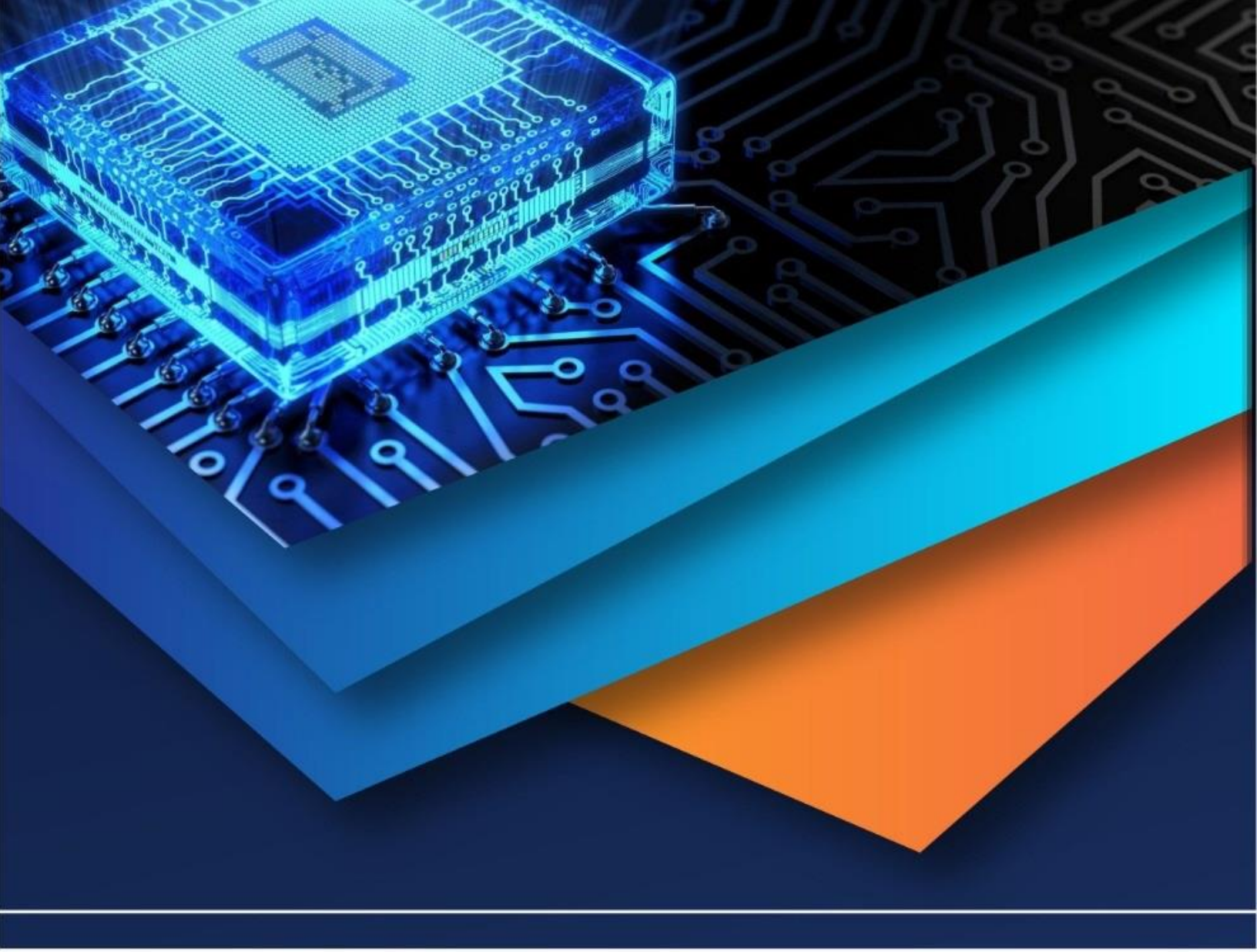
1) *Incorporation of Dynamic and Temporal Features:* To better capture the evolving nature of entity relationships, future research can incorporate temporal features into the knowledge graph. This will allow the system to analyse and anonymize data dynamically, addressing scenarios where relationships between entities change over time. This would enhance the applicability and scalability of the system to account for dynamic changes in the dataset.
2) *Expanding Domain Applicability:* While the current framework is demonstrated on a dataset with movie reviews, extending the methodology to sensitive domains like health- care, finance, or legal documents is a key area of future work. This would involve integrat- ing domain-specific knowledge bases, ontologies, and contextual nuances to enhance the effectiveness of privacy protection in these areas.
3) *Scalability and Performance Optimization:* With increasing dataset sizes, optimizing the system for scalability is crucial. Future work can focus on integrating distributed graph databases or parallel processing frameworks to handle millions of entities while maintaining performance and accuracy. The batch processing method explored in this research can be applied to larger datasets and methods to efficiently compute the risk scores can be tested.
4) *Real-Time Anonymization Capabilities:* Extending the system for real-time applications is a vital area for exploration. This could involve building frameworks capable of anonymiz- ing text streams, such as live reviews, social media posts, or chat logs, while maintaining computational efficiency and privacy guarantees. The proposed method can help prioritize the level of anonymization entities need to ensure sensitive information is hidden.
5) *Testing across Diverse Datasets and Attack Models:* Extending evaluations to more varied datasets, such as legal documents, emails, or patient records, and testing against a broader range of attack models, including linkage attacks and inference attacks, will vali- date the framework's effectiveness across different contexts.

REFERENCES

[1] Y. Sei, H. Okumura, T. Takenouchi, A. Ohsuga, Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness, IEEE transactions on dependable and secure computing 16 (4) (2017) 580–593.
[2] V. Alves, V. Rolla, J. Alveira, D. Pissarra, D. Pereira, I. Curioso, A. Carreiro, H. L. Cardoso, Anonymization through substitution: Words vs sentences, in: Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, 2024, pp. 85–90.
[3] M. Chiranjeevi, V. S. Dhuli, M. K. Enduri, K. Hajarathaiah, L. R. Cenkeramaddi, Quantifying node influence in networks: Isolating-betweenness centrality for improved ranking, IEEE Access.
[4] R. G. de Jong, M. P. van der Loo, F. W. Takes, A systematic comparison of measures for k-anonymity in networks, arXiv preprint arXiv:2407.02290.
[5] H. Wang, Revisiting local pagerank estimation on undirected graphs: Simple and optimal, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 3036–3044.
[6] T. Sasada, Y. Taenaka, Y. Kadobayashi, Anonymizing location information in unstructured text using knowledge graph, in: Proceedings of the 22nd International Conference on Information Integration and Web-based Applica- tions & Services, 2020, pp. 163–167.
[7] Y. J. Lee, K. H. Lee, What are the optimum quasi-identifiers to re-identify medical records?, in: 2018 20th Interna- tional Conference on Advanced Communication Technology (ICACT), IEEE, 2018, pp. 1025–1033.

[8] N. Torres, P. Olivares, De-anonymizing users across rating datasets via record linkage and quasi-identifier attacks, Data 9 (6) (2024) 75.
[9] L. Sweeney, k-anonymity: A model for protecting privacy, International journal of uncertainty, fuzziness and knowledge-based systems 10 (05) (2002) 557–570.
[10] Y. Yan, W. Wang, X. Hao, L. Zhang, Finding quasi-identifiers for k-anonymity model by the set of cut-vertex., Engineering Letters 26 (1).
[11] O. A. Ekle, W. Eberle, Dynamic pagerank with decay: A modified approach for node anomaly detection in evolving graph streams, in: The International FLAIRS Conference Proceedings, Vol. 37, 2024.
[12] A.-T. Hoang, B. Carminati, E. Ferrari, Protecting privacy in knowledge graphs with personalized anonymization, IEEE Transactions on Dependable and Secure Computing.
[13] I. B. C. Larbi, A. Burchardt, R. Roller, Clinical text anonymization, its influence on downstream nlp tasks and the risk of re-identification, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 2023, pp. 105–111.
[14] R. Yang, H. Liu, E. Marrese-Taylor, Q. Zeng, Y. H. Ke, W. Li, L. Cheng, Q. Chen, J. Caverlee, Y. Matsuo, et al., Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques, arXiv preprint arXiv:2403.05881.
[15] T. Yang, X. Zhu, I. Gurevych, Robust utility-preserving text anonymization based on large language models, arXiv preprint arXiv:2407.11770.
[16] S. R. Kodandaram, K. Honnappa, K. Soni, Masking private user information using natural language processing.
[17] P. Lison, I. Pila´n, D. Sa´nchez, M. Batet, L. Øvrelid, Anonymisation models for text data: State of the art, challenges and future directions, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4188–4203.
[18] X. Zhang, J.-W. van de Meent, B. C. Wallace, Disentangling representations of text by masking transformers, arXiv preprint arXiv:2104.07155.
[19] S. Sousa, R. Kern, How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing, Artificial Intelligence Review 56 (2) (2023) 1427–1492.
[20] A. Reeves, D. Ashenden, Understanding decision making in security operations centres: building the case for cyber deception technology, Frontiers in Psychology 14 (2023) 1165705.
[21] A. Javadpour, F. Ja'fari, T. Taleb, M. Shojafar, C. Benza¨ıd, A comprehensive survey on cyber deception techniques to improve honeypot performance, Computers & Security (2024) 103792.
[22] R. H. Weber, U. I. Heinrich, Anonymization, Springer Science & Business Media, 2012.
[23] I. E. Olatunji, J. Rauch, M. Katzensteiner, M. Khosla, A review of anonymization for healthcare data, Big data.
[24] D. K. Sharma, A. Sharma, A comparative analysis of web page ranking algorithms, International Journal on Com- puter Science and Engineering 2 (08) (2010) 2670–2676.
[25] F. Lamberti, A. Sanna, C. Demartini, A relation-based page rank algorithm for semantic web search engines, IEEE Transactions on Knowledge and Data Engineering 21 (1) (2008) 123–136.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)