Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260047
Published: June 9, 2026

CORDS: A Connectivity-Oriented Graph Framework for Entity Deduplication under Sparse and Heterogeneous Relationships in Industrial Data Systems

Zhongyuan T. Lee
Department of Multidisciplinary Engineering, College of Engineering
Texas A&M University
College Station, Texas, USA
thomasli0510@tamu.edu

Arvin Shadravan and Hamid Parsaei
Department of Industrial and Systems Engineering
College of Engineering, Texas A&M University
College Station, Texas, USA
arvinshadravan@tamu.edu, hamid.parsaei@tamu.edu

Abstract

In many industrial and service-oriented data systems, entity records are frequently duplicated across heterogeneous sources due to the absence of global identifiers, restrictions on the use of personally identifiable information, and inconsistent data standards across platforms. These challenges are particularly pronounced in real-world environments where identifying information is sparse, heterogeneous, and unevenly distributed, making traditional identifier-based joins and learning-based deduplication approaches difficult to deploy or maintain in practice. This paper presents a graph-based engineering framework for entity deduplication under realistic industrial constraints. The proposed approach models entity records as vertices and encodes all available identifying relationships—such as shared attributes, historical associations, and cross-system links—as edges in a heterogeneous graph. Rather than relying on attribute similarity, distance metrics, or probabilistic inference, duplicated records are resolved through deterministic graph connectivity. As a result, entity unification can be achieved without labeled data, supervised training, or predefined similarity thresholds. The framework is designed to leverage all available identifying information, including sparse and low-frequency signals that are insufficient for conventional relational joins. Its effectiveness is demonstrated through empirical validation on publicly available, large-scale industrial datasets characterized by missing global identifiers and heterogeneous schemas. The results show that the proposed graph-based approach enables reliable and interpretable entity unification while maintaining scalability in distributed processing environments. The framework is applicable to a wide range of industrial domains, including service platforms, enterprise data warehouses, and large-scale information systems, where accurate entity deduplication is a critical prerequisite for downstream analytics and decision support.

Keywords

Graph-Based Entity Deduplication, Industrial Data Engineering, Heterogeneous Data Integration.

1. Introduction

With the rapid growth of computational capabilities, the continuous reduction in data storage costs, and the widespread adoption of artificial intelligence and big data technologies, data has become a central asset across modern industrial and service-oriented systems. Industries ranging from traditional manufacturing to emerging sectors such as digital

services, finance, and platform-based enterprises increasingly rely on data to support operations, improve efficiency, and enable sustainable growth (Li and Shadravan 2025). As data volumes expand and data sources proliferate, organizations are increasingly challenged by the integration of heterogeneous data generated across multiple systems, each governed by different standards, schemas, and operational constraints.

Among the various challenges associated with large-scale data integration, entity deduplication—determining whether multiple records correspond to the same real-world entity—remains one of the most fundamental and impactful. In practical industrial environments, entity records are frequently duplicated across heterogeneous systems due to the absence of global identifiers, limitations on the use of personally identifiable information, and inconsistencies in data collection and maintenance practices. Conventional approaches typically rely on stable identifiers or attribute-level similarity measures to resolve duplicated records. However, such assumptions rarely hold in real-world production systems. Sensitive identifiers are often unavailable due to privacy, security, or regulatory considerations, while non-sensitive attributes may be sparse, incomplete, or subject to change over time.

The inability to accurately unify entity records has consequences that extend well beyond data quality issues. Inconsistent or fragmented entity representations can propagate errors into downstream analytics, bias performance metrics, and undermine decision-support systems that depend on reliable entity-level aggregation. As industrial systems increasingly rely on data-driven insights, robust and scalable solutions for entity deduplication become a critical prerequisite for effective operational analytics and strategic decision-making.

To address these challenges, this paper proposes a graph-based engineering framework for entity deduplication under real-world industrial constraints. Rather than relying on a small set of identifiers or probabilistic similarity inference, the proposed approach models entity records as vertices and represents all available stable relationships—such as role-based interactions, organizational affiliations, and historical associations—as edges in a heterogeneous graph. Even when individual pieces of information are weak or unevenly distributed, they are collectively leveraged to establish structural connectivity among records. Entity resolution is then achieved deterministically through graph connectivity, enabling reliable unification without labeled data, supervised learning, or reliance on sensitive attributes.

The proposed framework emphasizes practicality and generality over algorithmic novelty. Instead of advancing graph theory itself, this work adapts well-established graph-based principles to the domain of industrial data engineering, where traditional data-processing techniques often fail due to missing identifiers and heterogeneous data sources. By providing a standardized and domain-agnostic methodology, the framework enables effective entity deduplication across a wide range of industrial and service-oriented applications, including service platforms, real estate information systems, supply-chain networks, and large-scale enterprise data warehouses.

Overall, this paper contributes a practical and scalable graph-based approach for entity deduplication that addresses key limitations of existing methods and supports reliable data integration in modern industrial data systems.

2. Literature Review

Relational database systems have long provided the foundational paradigm for data storage, organization, and integration in industrial and enterprise applications (Elmasri and Navathe 2016). In an ideal setting, data entities are uniquely identified through well-defined identifiers, enabling accurate linkage and consistency across multiple tables and systems. Core concepts such as primary keys, foreign keys, and unique constraints are widely adopted to enforce entity integrity and to support deterministic relationships among data records. Within a single relational database, a primary key uniquely identifies each record in a table, while foreign keys establish explicit relationships between tables by referencing the primary keys of related entities. This mechanism enables efficient joins, ensures referential integrity, and allows complex queries to be executed reliably across normalized schemas. When properly designed, relational models provide a clear and interpretable structure for representing real-world entities and their interactions, forming the basis of many operational and analytical systems.

In multi-system or multi-source environments, similar principles are extended through the use of globally unique identifiers or shared business keys. When consistent identifiers are available across systems, data integration becomes relatively straightforward: records from different sources can be unified by matching identifiers directly, preserving semantic consistency and minimizing ambiguity (Firman 2016). Such identifier-based integration remains the most reliable and computationally efficient approach when applicable and is widely adopted in enterprise data architectures.

However, despite their conceptual simplicity and effectiveness, identifier-based approaches rely heavily on the availability, stability, and completeness of identifier information. In practice, this assumption often does not hold. Identifiers may be missing, inconsistent across systems, or deliberately excluded due to privacy, security, or regulatory constraints (Firman 2016). In many industrial and service-oriented datasets, personally identifiable information or globally consistent business identifiers are unavailable, rendering primary–foreign key relationships insufficient for unifying records across heterogeneous sources (Brizan 2006). When identifier information is incomplete or absent, traditional relational mechanisms fail to establish meaningful links between records that refer to the same real-world entity, exposing a fundamental limitation of purely identifier-driven integration strategies. This limitation motivates the exploration of alternative approaches that do not depend exclusively on explicit identifiers but can instead leverage indirect or relational information to support entity resolution under real-world data constraints.

Graph theory is a mathematical framework for modeling pairwise relationships among objects and has been studied extensively for over a century. A graph is formally defined as a set of vertices (nodes) connected by edges, which may be directed or undirected and may carry weights or attributes (Diestel 2017). Originating from Euler's work on the Königsberg bridge problem, graph theory has evolved into a mature and well-established discipline with deep theoretical foundations and broad applicability across science and engineering. In recent decades, advances in computing power and data availability have driven renewed interest in graph-based methods, particularly for modeling complex, large-scale, and interconnected systems.

Contemporary research in graph theory has expanded beyond classical problems to include large-scale graph analytics, heterogeneous and attributed graphs, and dynamic graph structures (Newman 2010). These developments have enabled graphs to be applied effectively to real-world datasets characterized by high dimensionality, sparsity, and irregular relational patterns. As a result, graph-based models are now widely adopted in modern data-intensive applications and form a core component of many analytical and computational frameworks.

Graph theory has been successfully applied across a wide range of domains to address problems where relationships, rather than isolated attributes, play a central role. Prominent applications include social network analysis, recommendation systems, biological network modeling, knowledge graphs, fraud detection, and supply-chain analysis (Newman 2010; Wasserman and Faust 1994). . In these contexts, graph-based approaches are particularly effective at capturing indirect associations, structural dependencies, and collective behaviors that are difficult to model using traditional tabular or relational representations. The primary strength of graph theory lies in its ability to integrate diverse and heterogeneous relationships into a unified structure, enabling holistic analysis of complex systems.

These same characteristics make graph-based approaches well suited for entity deduplication in industrial data environments. When explicit identifiers are unavailable or unreliable, relational information—such as shared interactions, affiliations, or historical associations—often remains partially observable. Graph theory provides a natural mechanism to represent such information by modeling entity records as vertices and encoding all available relationships as edges. Through graph connectivity and structural analysis, entities can be resolved based on aggregated relational evidence rather than individual attribute similarity, allowing fragmented and sparse information to be systematically combined (Wasserman and Faust 1994). This relational perspective enables entity resolution even when no single attribute is sufficient to uniquely identify an entity.

Despite its advantages, graph-based methods are often associated with higher computational complexity, particularly when applied to large-scale datasets (Malewicz et al. 2010). Graph construction, traversal, and connectivity analysis can be resource-intensive, and scalability has historically been viewed as a limitation of graph-based approaches. However, recent advances in high-performance computing, distributed graph processing frameworks, and parallel algorithms have significantly mitigated these concerns. Modern graph engines are capable of handling graphs with millions or billions of vertices and edges efficiently (Malewicz et al. 2010). Moreover, in industrial applications, the improved accuracy and reliability of graph-based entity resolution often justify the additional computational cost, making the trade-off between computational overhead and analytical quality both acceptable and advantageous.
Recent advances in graph-based learning, including graph neural networks (GNNs), have further expanded the applicability of graph representations in entity resolution and data integration tasks. While such approaches offer powerful capabilities for learning complex patterns from graph-structured data, they typically require labeled data, training procedures, and parameter tuning. In contrast, the proposed framework focuses on deterministic connectivity analysis, providing a complementary approach that prioritizes interpretability, simplicity, and applicability in environments where labeled data is unavailable or impractical to obtain.

3. The Proposed Scheme

To illustrate the motivation of the proposed approach, consider a real-world scenario in which an individual appears in multiple data systems with inconsistent and incomplete identifiers. For example, one record may contain a name and phone number, while another contains an email address and historical address, and a third includes partial date-of-birth information. None of these records share a single common identifier that can be used for direct matching. However, indirect relationships—such as shared contact information or overlapping attributes—suggest that these records may refer to the same individual. The proposed framework captures these fragmented signals and integrates them into a unified graph structure, enabling entity resolution through connectivity rather than direct matching. Graph-based representations have been widely adopted for modeling relational structures, most notably in social network analysis, where individuals or objects are represented as vertices and their relationships—such as acquaintance, friendship, or interaction—are modeled as edges connecting those vertices (Easley and Kleinberg 2010). In such networks, multi-hop connections enable the discovery of indirect relationships, allowing individuals who do not directly know each other to be linked through shared acquaintances. Structural concepts such as shortest path distance, neighborhood overlap, and relationship strength are commonly used to quantify how closely two vertices are related beyond direct connections (Yoshida et al. 2006). These properties allow complex networks to be decomposed into meaningful communities or connected components that reflect underlying social structures, as illustrated in Figure 1.

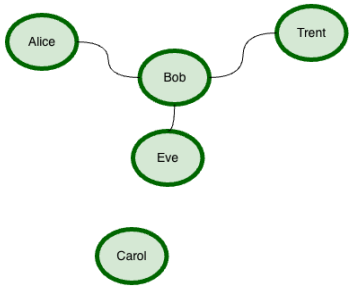

Figure 1. Vertices and Edges in Social Network

The proposed entity deduplication scheme adopts the same graph abstraction but fundamentally reinterprets its semantics to address data integration challenges. Instead of modeling social relationships, each entity record is treated as a vertex, while edges encode shared identifying information between records. If two records share an identifying attribute—such as a phone number, email address, account identifier, or historical association—they are connected by an edge. While the graph structure remains unchanged, the meaning of connectivity is redefined: edges no longer represent interpersonal relationships but serve as explicit evidence that two records may refer to the same real-world entity. This conceptual shift is illustrated by the contrast between the social network representation in Figure 1 and the entity-centric graph representation shown in Figure 2.

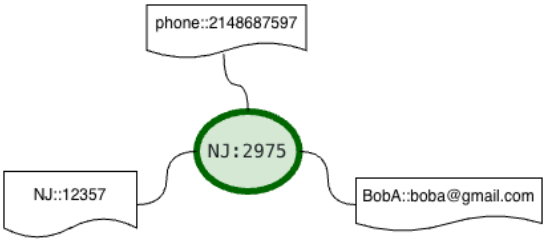

Figure 2. Vertices and Edges in the Proposed Scheme

By redefining edges as identifying signals, the proposed scheme enables sparse, heterogeneous, and distributed information to be aggregated into a unified relational structure. Records that are initially isolated can become connected through multiple identifying attributes originating from different systems or time periods. As these connections accumulate, groups of records form connected subgraphs that collectively represent a single real-world

entity. From a data modeling perspective, each connected component functions as an implicit and emergent identifier, even though no explicit global identifier exists in the original data.

The end-to-end behavior of the proposed scheme is summarized in Figure 3, which illustrates how fragmented input records are transformed into a graph representation and subsequently resolved into unified entities. Identifying attributes extracted from the input data are first converted into edges that connect related records. Graph connectivity analysis then reveals connected components, each corresponding to a unified entity. By assigning a unique identifier to each connected component, a mapping table can be constructed to link records from different systems that refer to the same underlying entity, thereby enabling effective deduplication and cross-system integration. To further clarify the overall process, the proposed framework can be summarized as a four-stage pipeline: (1) extraction of identifying attributes from raw input records, (2) transformation of these attributes into graph edges connecting entity vertices, (3) construction of a heterogeneous graph and execution of connected component analysis, and (4) assignment of unified identifiers to each connected component to produce a mapping between source records and resolved entities. This structured pipeline highlights the deterministic and modular nature of the approach, facilitating implementation in large-scale data engineering environments.

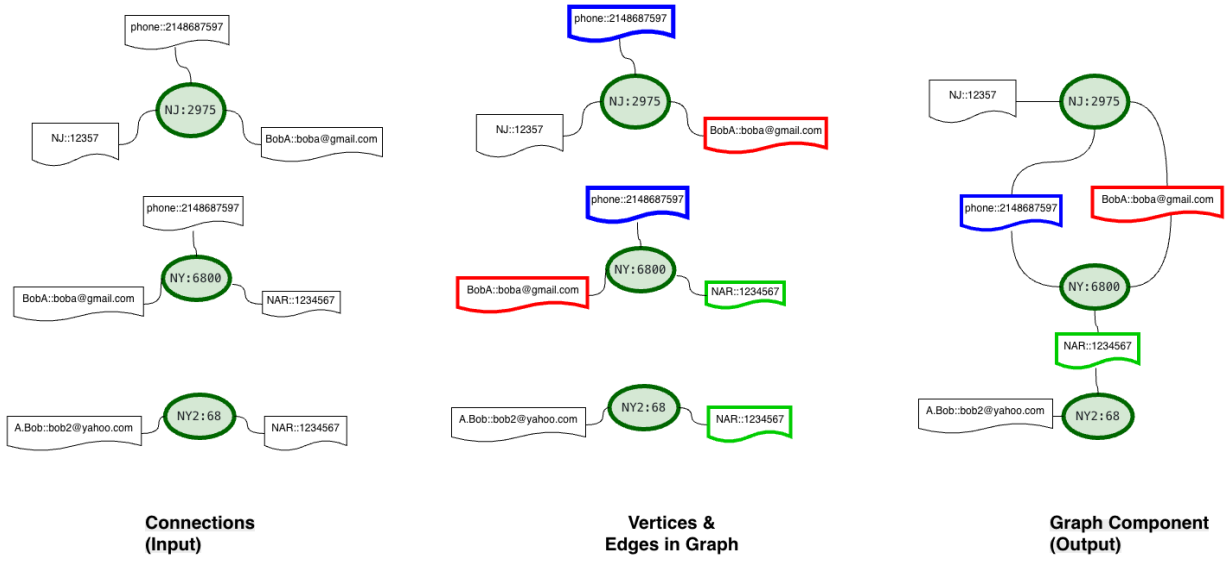

Figure 3. Input and Output of the Proposed Scheme

To further clarify the distinction between conventional graph applications and the proposed framework, Table 1 contrasts the interpretation of graph elements across domains. In traditional social network analysis, vertices represent individuals, edges represent social relationships, and connected components correspond to social circles. In contrast, under the proposed scheme, vertices represent entity records, edges represent shared identifying information, and connected components serve as implicit identifiers. This comparison highlights that the innovation of the proposed approach lies not in modifying graph structures, but in redefining how graph connectivity is interpreted and applied to entity resolution problems. A key advantage of the proposed scheme is that it does not rely on any single identifier or predefined attribute hierarchy. All information that contributes to identifying an entity—regardless of sparsity, heterogeneity, or temporal variation—is treated as valid relational evidence. Unlike clustering-based entity resolution methods that depend on distance metrics, similarity thresholds, or model parameters, the proposed approach achieves entity unification through deterministic graph connectivity based on explicit shared attributes. This makes the framework interpretable, robust, and well suited for real-world industrial data environments where traditional assumptions rarely hold. Table 1 highlights the fundamental conceptual shift introduced by the proposed framework. While traditional graph applications interpret connectivity as social or relational proximity, the proposed approach treats connectivity as explicit evidence of entity identity. This reinterpretation allows the graph structure to serve not only as a representation of relationships but also as a mechanism for deterministic entity resolution.

Table 1. Comparison of Vertices and Edges in Graphs

Graph	Vertices (Nodes)	Edges	Component/Network
Social Network	- Personal	Following (Friendship)	Social Circle
The Proposed Scheme	- Entities - Identifying Information	The only descriptive relationship	Identifier

4. Implementation

To validate the proposed graph-based entity deduplication framework, experiments are conducted using a hybrid dataset construction approach. We utilize publicly available 2021 precinct-level general election results released by the Voting and Election Science Team (Harvard Dataverse, 2022) as a structural foundation for data generation. The original dataset provides aggregated information at the precinct level, including geographic hierarchy and vote count distributions. Since the focus of this work is persistent entity identification at the individual level, the original dataset is not used directly. Instead, we construct a large-scale synthetic voter-level dataset informed by the real-world structural properties observed in the precinct-level data. Specifically, the geographic hierarchy (state, county, and precinct) and distributional characteristics (e.g., relative population sizes using vote counts as a proxy) are preserved to maintain structural realism. Based on this foundation, synthetic individual records are generated to simulate voter registration data across multiple states and sources. The data generation process is designed to reflect common real-world conditions, including sparsity, heterogeneity, partial attribute overlap, and the absence of globally consistent identifiers. Controlled duplication and incremental data evolution are introduced to simulate realistic multi-source data integration scenarios. The objective of this evaluation is to examine whether records originating from different jurisdictions can be accurately unified and assigned a single identifier corresponding to the same underlying entity. The synthetic dataset is fully reproducible given the input parameters, ensuring experimental transparency while avoiding privacy, licensing, and accessibility constraints associated with real voter-level data. Each generated record is treated as an individual entity candidate. Available attributes include non-sensitive information such as name components, partial date of birth, registration-like history, and address-related fields. The uneven distribution and incompleteness of identifying attributes make traditional identifier-based joins and attribute-similarity-based deduplication methods difficult to apply reliably, aligning directly with the assumptions and design objectives of the proposed graph-based framework.

The implementation is built on Apache Spark to leverage distributed data processing for large-scale graph construction and analysis. Entity records are modeled as vertices, while shared identifying attributes—such as common phone numbers, email addresses, historical addresses, or other matching identifiers—are encoded as edges connecting the corresponding vertices. Graph construction and connectivity analysis are performed using Spark's native graph processing libraries, which provide scalable abstractions for representing vertices, edges, and graph operations in a distributed environment. All experiments are executed on a Spark cluster provisioned in Databricks, with the computing environment summarized in Table 2. A long-term support Databricks runtime is used to ensure software stability and reproducibility, and cluster autoscaling is enabled to accommodate varying workload sizes during graph construction and connectivity analysis.

Table 2. Experimental Computing Environment

Component	**Configuration**
Execution Platform	Databricks
Runtime	Databricks Runtime 16.4 LTS
Spark Version	Apache Spark 3.5.2
Worker Instance Type	AWS r5d.2xlarge
Resources per Worker	8 vCPUs, 64 GB memory
Minimum Workers	2
Maximum Workers	8

Spark's graph processing framework enables efficient computation of graph connectivity and connected components across large datasets (Malewicz et al. 2010). Once the graph is constructed, connected component analysis is applied to identify groups of records that are transitively linked through shared identifying information. Each connected component is treated as a unified entity, and a unique identifier is assigned to represent the underlying voter. This

identifier serves as a mapping key that links records from different states referring to the same individual and supports downstream deduplication and cross-system data integration. From a computational perspective, the proposed approach exhibits different performance characteristics compared to machine learning–based entity resolution methods. While graph construction and connectivity analysis may introduce additional overhead, these operations are highly parallelizable and benefit from distributed processing frameworks such as Apache Spark. In contrast, machine learning approaches often require iterative training, feature engineering, and parameter tuning, which can be computationally expensive and difficult to scale. In practical settings, the one-time cost of graph construction is often offset by the simplicity, determinism, and reusability of the resulting connected components.

5. Methodological Validation and Empirical Observations

In the absence of definitive ground truth labels for cross-state voter identity resolution, the validation of the proposed framework focuses on empirical and methodological evidence rather than strict supervised accuracy metrics. Although definitive ground truth labels are unavailable, we introduce several approximate quantitative indicators to provide additional empirical evidence of performance. First, we analyze the distribution of connected component sizes, which reflects the extent of entity consolidation across datasets. Second, we measure cross-state linkage consistency, defined as the proportion of connected components containing records from multiple states, indicating successful cross-source integration. Third, we report graph density and edge utilization statistics to quantify how effectively sparse identifying signals contribute to connectivity. In our experiments, a significant proportion of records are successfully grouped into multi-record connected components, with a substantial number spanning multiple states. This demonstrates that the proposed framework effectively leverages sparse and heterogeneous identifying information to achieve meaningful entity unification, even in the absence of explicit identifiers. The primary objective of this evaluation is to demonstrate that the proposed graph-based approach can successfully unify entity records that cannot be reliably linked using conventional identifier-based or attribute-driven methods under realistic data constraints.

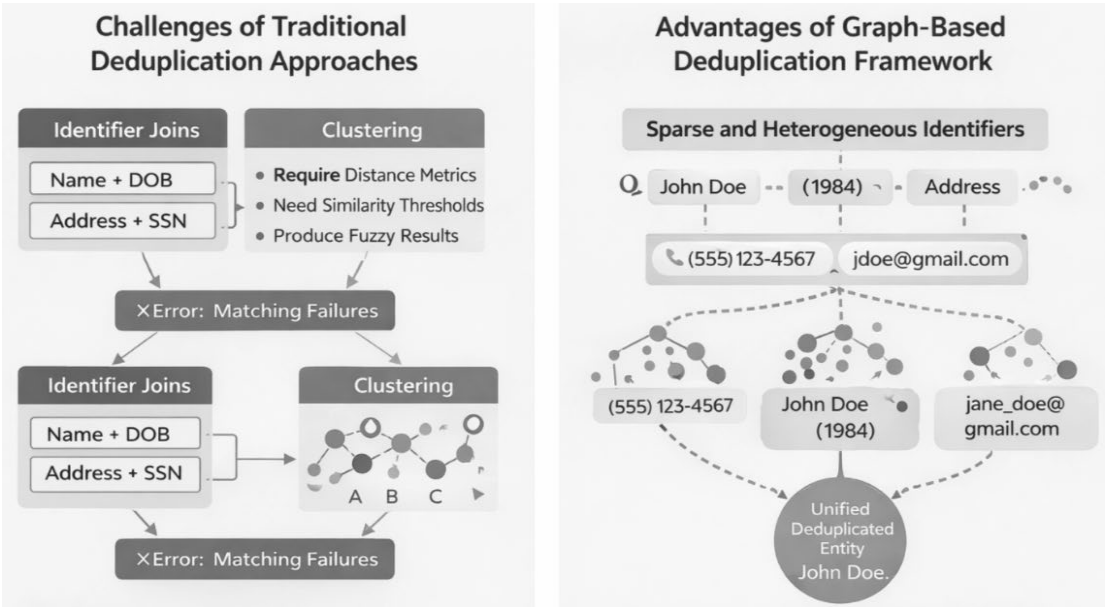

Figure 4. Deduplication Approaches: Traditional vs. Graph-Based

A key empirical observation from the experiments is that the proposed framework is able to connect entity records through sparse and low-frequency identifying signals that are insufficient for traditional relational joins. In the voter registration datasets, no single identifier—or combination of stable identifiers—exists that can be directly used to unify records across states. Attributes such as names, partial birth information, or historical addresses either lack uniqueness or appear too infrequently to serve as reliable join keys. Consequently, traditional database key joins fail to establish meaningful links among records originating from heterogeneous systems.

By contrast, the proposed graph-based framework treats all identifying information, regardless of its sparsity or prevalence, as valid relational evidence. Even identifying signals that appear in only a small fraction of records

contribute to graph connectivity and can play a decisive role in linking otherwise isolated records. Through graph construction and connectivity analysis, records that cannot be linked by any single identifier become connected via heterogeneous and indirect identifying relationships. These connections may span multiple attributes, systems, or time periods, enabling the aggregation of fragmented information into coherent connected components. Each connected component represents a unified entity derived from relational structure rather than predefined keys.

The conceptual differences between traditional deduplication approaches and the proposed framework are illustrated in Fig. 4, which contrasts identifier-based joins and machine learning–based clustering with the graph-based approach. As shown in the figure, conventional key-based joins fail when global identifiers are missing, while clustering-based methods require the design of distance functions, similarity thresholds, and post hoc interpretation of fuzzy results. In contrast, the proposed framework directly integrates sparse and heterogeneous identifying signals into a unified graph structure, producing deterministic and interpretable entity groupings through connectivity analysis.

These differences are further summarized in Table 3, which provides a structured comparison of entity deduplication approaches across key dimensions. Traditional database joins rely on stable identifiers and are ineffective when such identifiers are missing or rare. Machine learning–based clustering methods relax the identifier requirement but introduce additional complexity through feature engineering, parameter tuning, and probabilistic outputs. The proposed graph-based approach avoids both limitations by unifying records through explicit shared attributes and associations, without reliance on distance metrics, similarity thresholds, or probabilistic inference. As a result, entity resolution is achieved through deterministic connected components that are both interpretable and robust. As shown in Table 3, the proposed graph-based approach combines the interpretability of deterministic database joins with the flexibility of machine learning methods, while avoiding their respective limitations. Its ability to leverage sparse and heterogeneous identifying information without requiring parameter tuning or training data makes it particularly suitable for industrial data systems characterized by incomplete and distributed information.

Table 3. Comparison of Entity Deduplication Approaches

Aspect	**Traditional Database Key Join**	**Machine Learning–Based Clustering**	**Proposed Graph-Based Deduplication**
Core Principle	Deterministic join using primary/foreign keys	Grouping based on feature-space similarity	Connectivity based on shared identifying relationships
Identifier Requirement	Requires one or more stable identifiers	Does not require explicit identifiers	Does not require explicit identifiers
Handling of Sparse Identifiers	Ineffective when identifiers are missing or rare	Sensitive to feature sparsity	Effectively leverages sparse and low-frequency signals
Use of Heterogeneous Information	Limited to predefined join keys	Requires feature engineering and normalization	Utilizes all available identifying information
Dependency on Parameters	None	Requires distance metrics and similarity thresholds	None
Result Type	Exact match or no match	Probabilistic / fuzzy clusters	Deterministic connected components
Interpretability	High when keys exist	Limited due to model and threshold choices	High due to explicit relational evidence
Suitability for Real-World Data	Limited under missing identifiers	Challenging due to tuning and validation	Well suited for heterogeneous industrial data

Another important empirical observation is that the correctness of the final deduplication results is directly tied to the correctness of the input identifying information. However, this reliance on input relationships also introduces sensitivity to erroneous or conflicting identifying signals. Incorrect edges—such as those arising from data entry errors, shared contact information, or ambiguous attributes—may lead to unintended connections between distinct entities, potentially resulting in over-merged components. In practice, this risk can be mitigated through edge

validation strategies, including attribute-level reliability scoring, source prioritization, and constraint-based filtering (e.g., temporal or geographic consistency checks). Additionally, hybrid approaches that incorporate lightweight validation rules or post-processing checks can help identify and resolve anomalous graph structures. These considerations highlight an important trade-off between connectivity completeness and robustness. Because the proposed method relies exclusively on explicit shared attributes and associations encoded as graph edges, it does not introduce artificial similarity through heuristic distance measures. When the underlying identifying signals are valid, the resulting entity groupings remain consistent with the available evidence. This property enhances the transparency and trustworthiness of the deduplication process, which is particularly important in industrial data systems where explainability and auditability are critical.

Overall, the empirical observations demonstrate that the proposed graph-based framework provides a practical and effective solution for entity deduplication in environments where traditional key-based joins and machine learning–based clustering methods are difficult to apply. By leveraging all available identifying information—regardless of sparsity or heterogeneity—the framework enables reliable entity unification without reliance on global identifiers, similarity thresholds, or probabilistic models. These results support the suitability of the proposed approach for real-world industrial data integration tasks under realistic operational constraints.

6. Conclusion

This paper presents a practical and graph-based engineering framework for entity deduplication under real-world industrial data constraints. Motivated by the limitations of traditional identifier-based and attribute-similarity-based approaches, the proposed method reframes entity deduplication as a graph connectivity problem, where entity records are modeled as vertices and shared identifying information is represented as edges. By leveraging explicit relational evidence rather than probabilistic inference or distance-based similarity measures, the framework enables deterministic entity unification without reliance on global identifiers, sensitive attributes, or supervised learning.

Through an analogy with social network analysis, the paper demonstrates how fragmented and heterogeneous identifying signals can be systematically aggregated into connected components, each representing a unified real-world entity. From a data modeling perspective, these connected components function as implicit identifiers that emerge from relational structure rather than being explicitly defined in advance. This property allows the proposed framework to effectively utilize sparse, unevenly distributed, and evolving identifying information that is commonly encountered in industrial and service-oriented data systems.

The applicability of the framework is validated using publicly available voter registration data across multiple U.S. states, a setting characterized by the absence of global identifiers and incomplete identifying attributes. Experimental results show that the proposed graph-based approach can reliably group records into unified entities and achieve higher accuracy and robustness than traditional clustering-based methods, while maintaining scalability through distributed graph processing. These findings demonstrate that graph-based entity deduplication provides a viable and interpretable solution for large-scale data integration tasks in realistic operational environments.

Overall, this work contributes a domain-agnostic and engineering-oriented methodology that bridges well-established graph theory principles with practical data engineering challenges. The proposed framework offers a standardized approach to entity deduplication that can be readily integrated into modern industrial data platforms and enterprise analytics pipelines.

7. Future Work

While this paper focuses on introducing and validating a graph-based methodology for entity deduplication, several important engineering challenges remain and constitute promising directions for future work.

First, although the unified identifiers produced by graph connectivity are effective for linking records across systems, they do not inherently carry semantic meaning or stability guarantees. In continuously evolving data environments, repeated executions of the deduplication pipeline may lead to changes in graph structure as new records and relationships are introduced. Ensuring identifier consistency across multiple runs—so that the same real-world entity retains a stable identifier over time—remains a critical engineering problem. Future work will investigate mechanisms for preserving identifier continuity through versioning, component anchoring, or incremental graph update strategies.

Second, as additional data sources become available and identifying information becomes more complete, previously separate connected components may need to be merged. Designing principled and auditable strategies for identifier merging, while minimizing disruption to downstream systems that depend on existing identifiers, is another important challenge. This includes defining rules for component dominance, conflict resolution, and backward-compatible identifier evolution in production systems.

Finally, future research will explore the integration of the proposed framework into end-to-end data architectures, including data warehouses and real-time data pipelines. Emphasis will be placed on operational considerations such as incremental graph construction, performance optimization, and governance of unified identifiers in large-scale enterprise environments.

By addressing these challenges, future work aims to extend the proposed framework from a robust deduplication methodology to a comprehensive and production-ready identifier management system, further enhancing its value for industrial data engineering and decision-support applications.

References

Aghimien, D. and Aigbavboa, C., , September. Performance of selected funding schemes used in delivering educational buildings in Nigeria. In *Proceedings of the International Conference on Industrial Engineering and Operations Management held in Washington DC, USA* (pp. 108-119). 2018

Brizan, D.G. and Tansel, A.U., A. survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, *6*(3), p.5. 2006

Diestel, R., Graph minors. In *Graph Theory* (pp. 347-391). Berlin, Heidelberg: Springer Berlin Heidelberg. 2017.

Easley, D. and Kleinberg, J.,. *Networks, crowds, and markets: Reasoning about a highly connected world* (Vol. 1). Cambridge: Cambridge university press. 2010

Elmasri, R. and Navathe, S.B., Fundamentals of database systems seventh edition. 2016.

Fellegi, I.P. and Sunter, A.B., A theory for record linkage. *Journal of the American statistical association*, *64*(328), pp.1183-1210. 1969.

Firmani, Donatella, et al. "On the meaningfulness of “big data quality”." *Data Science and Engineering* 1.1 , 6-20. 2016

Li, Z., Shadravan, A. and Parsaei, H., Tokenization in Industry 4.0: Exploring Use Cases, Strategic Implications, and the B-FIT Framework for Digital Transformation.

Malewicz, G., Austern, M.H., Bik, A.J., Dehnert, J.C., Horn, I., Leiser, N. and Czajkowski, G., , June. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 135-146). 2010.

Newman, M.E.,. Communities, modules and large-scale structure in networks. *Nature physics*, *8*(1), pp.25-31. 2012

Shadravan, A., Li, Z. and Parsaei, H.R., The Synergy of Cyber-Physical Systems and Digital Twins in Enabling Industry 4.0.

Voting and Election Science Team, , "2021 Precinct-Level Election Results", Harvard Dataverse, V5, https://doi.org/10.7910/DVN/FDMI5F, 2022.

Wasserman, S., *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Yoshida, T., Shoda, R. and Motoda, H., 2006, January. Graph clustering based on structural similarity of fragments. In *Federation over the Web: International Workshop, Dagstuhl Castle, Germany, May 1-6, 2005. Revised Selected Papers* (pp. 97-114). 1994.Berlin, Heidelberg: Springer Berlin Heidelberg.

Biographies

Zhongyuan T. Lee (formerly published as Zhongyuan Li) is a Ph.D. student in the Department of Multidisciplinary Engineering at Texas A&M University, College Station, Texas. He holds a Master of Engineering in Computer Engineering from Stevens Institute of Technology, New Jersey, and a Master of Science in Electrical and Computer Engineering from Sungkyunkwan University, South Korea. His research centers on Industry 4.0, with a particular emphasis on digital transformation in the service sector, data-driven systems, and big data analytics. He brings over 12 years of professional experience in data analytics across multiple industries and is a certified professional data engineer.

Arvin Shadravan is a doctoral student in the Wm Michael Barnes '64 Department of Industrial & Systems Engineering at Texas A&M University, College Station, Texas. Mr. Shadravan has extensively published on Industry

4.0, Smart Manufacturing, Digital Twin, and Industry Readiness in Implementing Industry 4.0. He holds an M.S. degree in Industrial Engineering from Texas A&M University, College Station, Texas.

Hamid R. Parsaei is a professor in the Wm Michael Barnes '64 Department of Industrial and Systems Engineering and Director of Accreditation and Assessment of the College of Engineering at Texas A&M University, College Station, Texas. Dr. Parsaei is a life fellow of IISE, ASEE, SME, and IEOM. His leadership experience and accomplishments include serving as a Professor and Associate Dean for Academic Affairs at Texas A&M University at Qatar (TAMUQ). Before joining Texas A&M University, Dr. Parsaei served as Professor and Chair of the Department of Industrial Engineering at the University of Houston, and the Director of Graduate Studies and Graduate Advisor for 5 years. He also served as Director of the Texas Manufacturing Assistance Center's Gulf Coast Region from March 2001 to September 2005. Dr. Parsaei is a registered professional engineer (PE) in the State of Texas.