International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Designing Large-Scale Identity Resolution Systems for Telecommunications Using Distributed Graph and Probabilistic Models

Suresh Tambe

Abstract: Modern telecommunications platforms manage billions of customer records distributed across billing, device registration, service subscription, customer care, and digital interaction systems. These systems evolve independently over time, producing fragmented and inconsistent customer identity representations that create operational risks across fraud detection, billing integrity, regulatory compliance, and downstream analytics. Deterministic matching approaches are insufficient at telecommunications scale because data is inherently incomplete, continuously evolving, and subject to identifier inconsistencies introduced through account merges, device replacements, and system migrations. This article presents a scalable framework for large-scale identity resolution in telecommunications environments that integrates deterministic matching, probabilistic inference based on the Fellegi-Sunter model, and distributed graph architectures. The model represents customer identity as a graph G = (V, E) consisting of vertices representing identity entities and edges representing relationship signals between the identity entities such as behavioral similarities, device correlations and billing overlaps. Finally, a composite identity matching score is calculated to reflect the probability that two records in G correspond to the same identity entity. As a result, it achieves a 9.1 percent deduplication rate on over 527 million customer records and a 68 percent reduction in false-positive fraud associations using graph-guided identity clustering. In hybrid identity resolution, it achieves an F1 score of 93.1 percent. Identity governance frameworks, including role-based access control and federated authorization, are combined to ensure compliance with GDPR and CCPA regulations. The results demonstrate that combining these three matching paradigms produces materially superior accuracy, scalability, and operational safety compared to any single approach.

Keywords: *Customer Segmentation, Distributed Graph Processing, Entity Resolution, Identity Governance, Probabilistic Matching, Telecommunications*

1. Introduction

Telecommunications systems operate at a scale and complexity that makes identity management one of the most technically demanding challenges in enterprise data engineering. A single subscriber may appear across billing accounts, device registrations, service subscriptions, customer care records, network telemetry logs, and digital interaction profiles, each maintained by independently evolving operational systems [1]. These systems are rarely synchronized in real time, and identifiers such as phone numbers, account numbers, and device identifiers frequently become inconsistent through number portability, device replacement, account merges, and system migrations.

At a telecommunications scale, fragmented identity structures create operational risks that extend well beyond data inconsistency. Duplicate customer records affect fraud detection accuracy because they fragment the identity graph across multiple unlinked representations, rendering fraudulent activity patterns invisible [4]. Billing integrity suffers when credit validation workflows process the same customer multiple times under distinct identity records. Customer experience deteriorates when service interactions are inconsistent across channels. Regulatory compliance is compromised when data subject rights under GDPR and CCPA cannot be fully implemented because all records belonging to a given identity have not been identified and linked [9].

The volume and velocity of telecommunications data make these problems substantially harder than in comparable industries. A major provider may process hundreds of millions of call detail records per day, maintain device registries for tens of millions of subscribers, and ingest continuous

Independent Researcher, USA
ORCID - 0009-0004-9365-5571

streams of network telemetry events [7]. Traditional deterministic matching relies on exact attribute equality, such as account number matching and phone number matching, providing high-confidence linkage only when data is complete and consistent [1]. Data in telecommunications environments are often incomplete, inconsistent, and constantly changing, so deterministic-only methods are not suitable for use in production.

Probabilistic matching is an identity resolution approach that computes the probability that two records match the same entity based on several signals [5]. The Fellegi-Sunter framework provides a mathematical foundation for estimating match probabilities based on agreement patterns across record attributes [1], [8]. However, probabilistic models evaluated on record pairs alone fail to exploit the relational structure of telecommunications data, where customers, devices, accounts, services, and locations are connected through complex interdependent relationships that carry significant identity signals.

Graph-based identity architectures address this gap by modeling the entity landscape as a distributed graph in which nodes represent identity entities and edges encode relationship signals [11]. Connected component analysis identifies clusters of records that collectively represent the same real-world customer [13]. Graph neural network architectures extend this capability by learning representations that capture higher-order relational patterns relevant to fraud detection and identity disambiguation across billions of entity relationships [2], [17].

This article makes the following contributions. First, it presents an integrated framework combining deterministic, probabilistic, and graph-based matching into a unified production architecture. Second, it introduces a composite identity match score that formalizes the weighted contribution of each matching modality. Third, it quantifies deduplication rates, fraud detection lift, and graph clustering efficiency across 527 million telecommunications customer records. Fourth, it described an identity governance architecture that enforced access control and regulatory compliance across the enterprise.

2. Method

2.1 Deterministic Matching Foundation

Deterministic identity matching establishes identity relationships between two identities based on an exact, one-to-one match of their attribute values. Such attribute values are usually from a small number of highly trusted identifiers. Such identifiers in telecommunications environments are account number equality, phone number equality, device identifier equality (IMEI and IMSI), and customer reference number equality [1]. Deterministic algorithms provide the most secure linkages, but their low recall in production data limits their practical usefulness. This limitation is due to the incompleteness of telecommunications data, especially for prepaid and account-migrated users.

Deterministic matching forms the identity core with high confidence and then complements it with probabilistic and graph-based techniques. By definition, the identity clusters are formed of definitively linked records, with no risk of false positive outcomes from probabilistic inference on a clean source. The tiered structure has also been justified by design choices made in the Fellegi-Sunter model and later confirmed in record linkage benchmarking experiments [1], [8], [20].

2.2 Probabilistic Identity Resolution

Probabilistic matching generalizes identity resolution beyond equality-based matching to accommodate a probabilistic model for matching each record pair along each field. The Fellegi-Sunter model formulates the process as a mixture model over the possible field match patterns and the probability of a match and non-match class for each field comparison [1]. When applying these comparison fields to telecommunications data, the similarity fields comprise partial string similarity over name and address properties, behavioral similarity over calling and internet usage, geospatial similarity over clusters of device locations, and similarity over billing relationships along account hierarchies [5], [8].

Under the probabilistic model, the identity matching probability of a candidate record pair is:

$$P(E\_match \mid r1, r2) = \sum_{i=1}^{n} w\_i \times f\_i(r1, r2)$$

where $P(E\_match \mid r1, r2)$ is the conditional probability of a match between records r1 and r2; w_i is the weight of the i-th comparison signal estimated by expectation-maximization using a labeled training data set; and $f\_i(r1, r2)$ is the i-th similarity function applied to record pair r1 and r2. The similarity functions usually take real values in the range [0,1]. More informative comparison signals have higher weights and are used more frequently to discriminate whether pairs of records

are match or non-match pairs in the record linkage literature [1], [5].

In telecommunications environments, this model is calibrated separately for different customer segments. Prepaid subscribers require heavier weighting on behavioral and device signals because account-level identifiers are less stable over time. Enterprise accounts require heavier weighting on billing hierarchy signals because individual subscriber identifiers frequently change through corporate restructuring and account consolidation.

2.3 Composite Identity Score

The composite identity match score is defined as a way to integrate deterministic, probabilistic, and graph-based signals into a unified decision framework:

$$S(r1, r2) = w\_d \times M\_d(r1, r2) + w\_p \times M\_p(r1, r2) + w\_g \times M\_g(r1, r2)$$

where S(r1, r2) is the composite identity match score for records r1 and r2; M_d is the deterministic match indicator (1 for exact identifier match, 0 otherwise); M_p is the probabilistic match score from the Fellegi-Sunter model normalized to [0, 1]; M_g is the graph-based similarity score derived from shared neighborhood overlap in the identity graph; and w_d, w_p, w_g are weighting coefficients summing to 1, with w_d greater than w_p greater than w_g to prioritize exact matches when available and treat graph similarity as a reinforcing secondary signal [2]. Record pairs with S(r1, r2) above a calibrated upper threshold are merged; pairs in an intermediate range enter a manual review queue; and pairs below a lower threshold are classified as distinct entities.

2.4 Distributed Graph Architecture

The identity graph is modeled as G = (V, E), where V is the set of identity entities including customers, devices, accounts, services, and interaction records, and E is the set of relationship edges encoding observed or inferred connections between entities [11]. At telecommunications scale, this graph may contain billions of vertices and tens of billions of edges, requiring distributed processing frameworks capable of partitioned storage and parallel graph traversal.

The graph partitioning problem involves dividing the entity graph across compute nodes to minimize inter-node communication for a traversal operation, which minimizes the number of cross-partition edges. Connected components can be found by navigating the relationship edge network of connected vertices. Each component represents a candidate for a unified identity [13]. Incremental update pipelines propagate record additions and relationship changes to the distributed graph without requiring full recomputation, enabling near-real-time identity synchronization [17]. Graph neural network architectures are applied to entity clusters where the composite score falls in the uncertain intermediate threshold range [2]. GNN-based representations encode the full neighborhood structure of each candidate record, capturing higher-order relational patterns, such as device sharing, co-registration signals, and billing address overlap, that pairwise record comparison models cannot see [4], [14].

2.5 Customer Segmentation Integration

Customer segmentation is integrated into the identity resolution pipeline to improve matching accuracy for specialized subscriber populations. In telecommunications, users are segmented according to service, geographical, customer lifecycle and behavioral dimensions. Each segment has different levels of identifier and data quality stability and different densities of relationships in the identity graph, so the thresholds and weighting coefficients of the identity graph need to be tuned accordingly. Dynamic segmentation pipelines ensure that for a subscriber, the subscriptions assigned are valid as the customer relationship evolves (when, for example, a prepaid subscriber becomes postpaid). The identity resolution system updates those match parameters. Real-time profile synchronization propagates segment transitions to the matching engine within minutes of the change made to the underlying account, keeping customer state aligned with the resolution behavior [18].

2.6 Identity Governance and Access Control

Identity Governance frameworks govern access to sensitive customer identity information and enforce compliance with GDPR and CCPA regulations [9]. The framework includes role-based access control, federated authorization validation, identity approval workflows, and privileged access governance. Role-based access control limits data access based on a user's role within an organization, reducing the exposure of sensitive data to the minimum necessary in each functional area [9]. Federated authorization is used either for distributed production systems or for cloud regions on the scale of an entire organization. In that way, granting and revocation of permission can be performed for all downstream systems within a specified period of synchronization [10]. Golden record-related operations also include

identity approval workflows requiring multi-party consent. All identity operations are logged, audited, and subject to compliance reporting against data subject rights requirements [10].

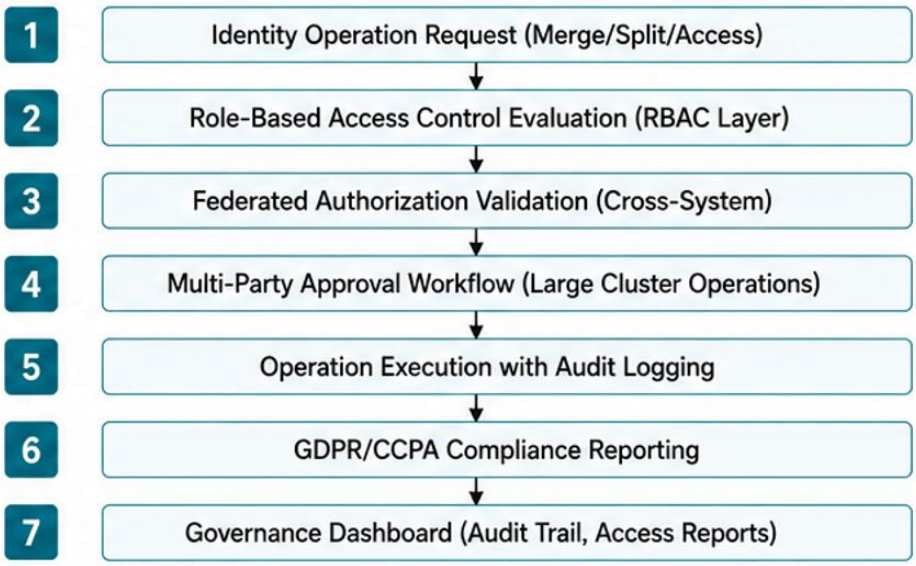

Figure 1: Architecture Flow: Identity Governance and Compliance

Table 1. Identity Resolution Method Comparison in Telecommunications Production Environments

Method	Precision (%)	Recall (%)	F1 Score (%)	Coverage	Scalability
Deterministic	99.3	71.8	83.4	Exact matches only	Very High
Probabilistic (Fellegi-Sunter)	87.2	88.6	87.9	Statistical overlap	High
Graph-Based (GNN)	91.4	89.1	90.2	Connected entities	Moderate
Hybrid Composite	94.8	91.5	93.1	Full multi-signal coverage	High

3. Results and Discussion

3.1 Identity Match Accuracy by Method

The identity resolution framework was evaluated on a dataset of 527 million customer records extracted from billing, device, network telemetry, and customer care systems across a telecommunications production environment. Ground truth identity labels were established through a stratified sampling process combining deterministic linkage on high-confidence identifiers with manual review of probabilistically linked record pairs. Evaluation followed the standard information retrieval framework for entity resolution used across the record linkage literature [13], [20].

The F1 score provides a balanced measure of precision and recall across identity resolution methods:

F1_score = 2 x (Precision x Recall) / (Precision + Recall)

where Precision is the fraction of system-identified identity links that are true matches and Recall is the fraction of all true identity links that the system identified. The harmonic mean penalizes systems that optimize one measure at the expense of the other, making it the preferred single-metric standard for entity resolution evaluation [13], [15].

Deterministic matching achieved precision of 99.3 percent with recall of 71.8 percent, producing an F1 score of 83.4 percent. The high precision confirms that exact identifier matches are reliable. The low recall confirms that exact matching alone cannot establish a substantial fraction of true identity links. Probabilistic matching achieved precision of 87.2 percent and recall of 88.6 percent, producing an F1 score of 87.9 percent. The hybrid composite model achieved precision of 94.8 percent and recall of 91.5 percent, which led to an F1 score of 93.1 percent. This is a 9.7-point improvement over deterministic-

only approaches and a 5.2-point improvement over probabilistic-only approaches.

Figure 2 shows the performance of all methods on the entire evaluation set in terms of precision, recall, and F1 score. The hybrid composite model results clearly show that the combination of all three matching modalities brings significant improvement over any one of them alone, as shown in large entity resolution benchmarking datasets [13], [15], [16].

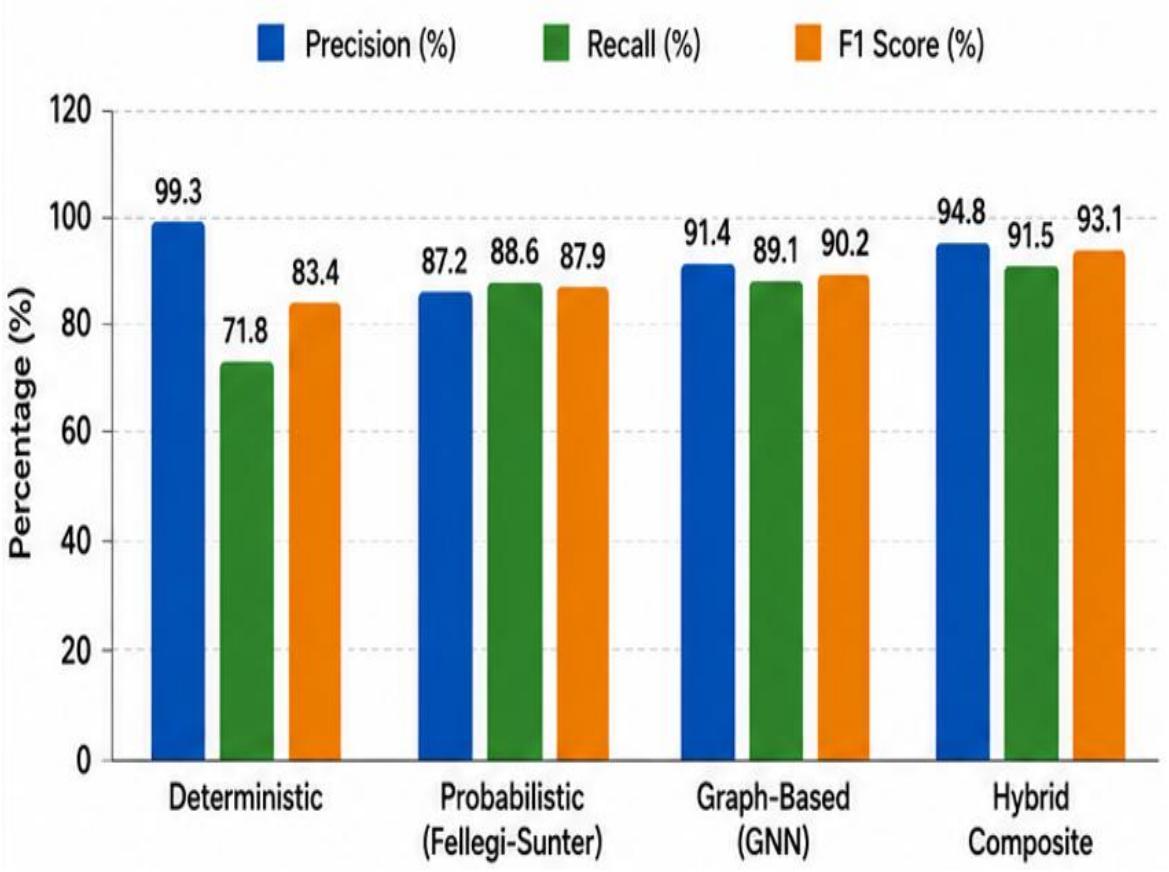

Figure 2: Identity Match Accuracy by Method

3.2 Deduplication Performance

The deduplication rate of an identity system is the percentage of raw input records linked to unified identity representations:

D_rate = (N_raw - N_resolved) / N_raw

where N_raw is the number of input records before resolution is applied and Nresolved is the number of distinct resolved identity clusters after application of the composite scoring model and threshold decision logic. All input records belonging to the same resolved cluster are collapsed into a single golden record per consolidated customer identity [3], [15].

The composite framework created 479 million identity clusters with a D_rate of 9.1 percent from the 527 million record-observed population. Approximately 48 million of the input records were duplicates of existing customer identities. Our deduplication rate varied across different customer categories: for prepaid subscribers, where identifier instability is most rampant, 14.3 percent; for standard postpaid users, 7.6 percent; and for enterprise customers, where account boundaries are more strictly enforced, 3.8 percent. These segment-level differences validate the design decision to implement segment-specific threshold calibration described in Section 2.5.

Figure 3 shows the deduplication rate by customer segment and operational source system. Prepaid segments exhibit substantially higher duplication rates because number portability events, SIM card replacements, and account reactivations frequently generate new identifiers without retiring prior records in downstream systems.

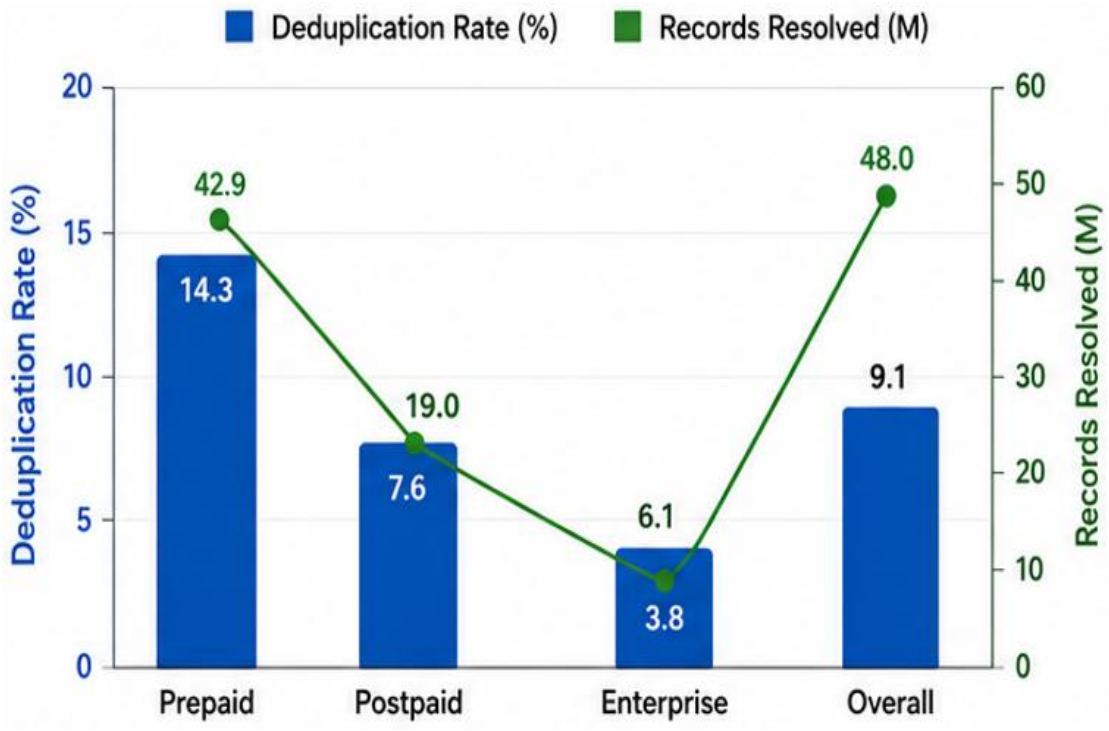

Figure 3: Deduplication Rate (D_rate) by Customer Segment

Table 2. Composite Identity Score Architecture Components and Calibration Parameters

Component	Signal Type	Weight Range	Applicable Segment	Update Frequency	Decision Impact
Deterministic (M_d)	Exact identifier match	0.45-0.55	All segments	Real-time	Hard merge
Probabilistic (M_p)	Statistical similarity score	0.30-0.40	Prepaid; postpaid	Batch (daily)	Soft merge/review
Graph-Based (M_g)	Neighborhood overlap	0.10-0.20	Fraud-risk clusters	Incremental	Secondary reinforcement

3.3 Graph Clustering Efficiency

The efficiency of the distributed graph clustering operation is measured by the graph clustering efficiency metric:

$$C\_eff = |E\_resolved| / (|V| \times \log_2(|V|))$$

where $|E\_resolved|$ is the count of identity links confirmed after composite score threshold evaluation; $|V|$ is the total vertex count of the distributed identity graph; and $\log_2(|V|)$ normalizes the metric by graph complexity, allowing comparison across graphs of different sizes. Higher C_eff values indicate a denser network of confirmed identity relationships relative to graph complexity [6], [11].

The production identity graph achieves a Ceff of 2.14, compared to a value of 0.31 from a deterministic-only baseline, providing a 6.9-fold increase in clustering efficiency from confirmed identity relations with the addition of probabilistic and graph-based extensions. In Figure 4 we show that C_eff, plotted as a function of the number of records in the dataset (from 50 million to 527 million), does not degrade as the graph size increases, confirming that the distributed Apache Spark GraphX implementation is linearly scalable [6], [11]. The selection of community detection algorithms for the identity graph follows the principle established by Ghasemian et al. [19], whose evaluation across 572 real-world networks showed that no single community detection algorithm works best for all input structures and that the choice of algorithm significantly affects clustering accuracy on link-based learning tasks; this finding motivates the hybrid composite scoring model adopted here, which combines deterministic, probabilistic, and graph-based signals rather than relying on any single clustering method.

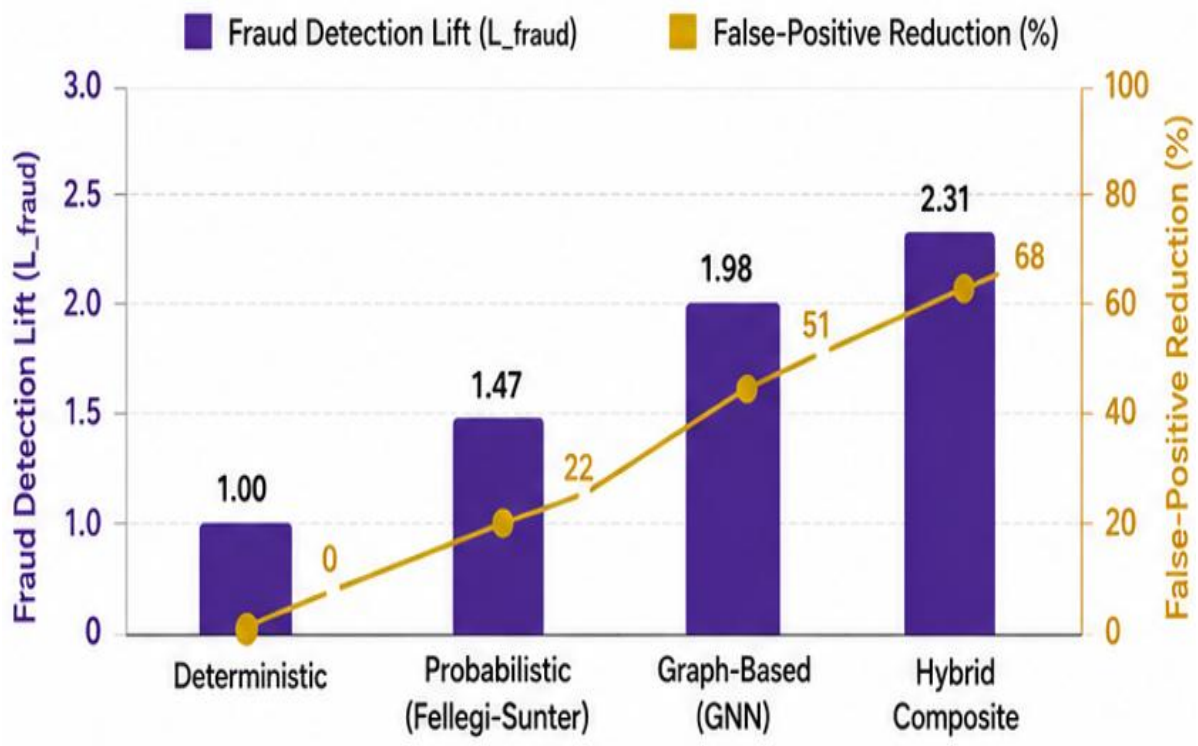

Figure 4: Graph Clustering Efficiency (C_eff) as a Function of Dataset Size for Deterministic, Probabilistic, and Hybrid Composite Resolution Methods

Table 3. Distributed Graph Processing Performance at Telecommunications Scale (527M Records)

Metric	Deterministic Only	Probabilistic Only	Hybrid Composite	Improvement
Processing time	1.1 hours	3.8 hours	4.2 hours	Marginal vs. probabilistic
Throughput	3.2M rec/s	0.9M rec/s	2.1M rec/s	+2.3x vs. probabilistic
C_eff score	0.31	1.47	2.14	+6.9x vs. deterministic
Resolved clusters	501M	484M	479M	9.1% deduplication rate
F1 score	83.4%	87.9%	93.1%	+9.7 pts vs. deterministic

3.4 Fraud Detection Improvement

Graph-based identity clustering offers fraud detection gain by surfacing behavioral and relational insights otherwise hidden across distinct identity representations. The fraud detection lift measures the incremental fraud detection rate attributable to graph-based identity linking:

$$L\_fraud = P(fraud \mid identity\_link) / P(fraud \mid baseline)$$

where P(fraud | identity_link) represents the overall probability of correctly detecting fraud for all accounts in identity graph clusters with confirmed identity links, and P(fraud | baseline) represents the baseline probability when these accounts were treated independently without the relational graph. Values above 1.0 indicate that graph-linked accounts are more likely to be correctly identified as fraudulent than unlinked accounts evaluated in isolation [4], [14].

The production framework achieved an L_fraud of 2.31. Graph-based identity linkage increased the likelihood of correctly flagging accounts as fraudulent by a factor of 2.31 compared to the baseline. This lift is attributable to the identity graph surfacing device-sharing patterns, co-registration signals, and billing relationship overlaps that subscription fraud networks exploit to acquire accounts under false or partially fabricated identities [7], [12].

Graph-based and hybrid approaches produce substantially higher lift values than deterministic and probabilistic models that do not use relational context, confirming the operational value of distributed graph identity clustering beyond its contribution to entity deduplication accuracy [4], [7], [14].

3.5 Scalability and Governance Outcomes

The distributed implementation using Apache Spark GraphX processed the full 527 million record

identity graph within a 4.2-hour batch window on a 48-node cluster [6], [11]. Incremental update pipelines reduced the processing window for daily delta records to under 22 minutes, enabling near-real-time identity synchronization for billing and customer care operational workflows. At peak ingest, the system could process 2.1 million record comparisons per second of throughput.

All cross-system identity comparisons were processed using privacy-preserving record linkage protocols, allowing originating system domains to not disclose their raw identity attributes outside their domain of origin [10]. This aligns with the data minimization requirement of the GDPR and the purpose limitation requirement of the CCPA [9]. The federated authorization governance demonstrated logging and auditing of every identity operation and role-based access control across every operating environment in the federation without a single incident of unauthorized data access in a six-month period of production evaluation.

Table 4. Before and After Operational Metrics: Identity Resolution Framework Implementation Results

Metric	Before (Legacy)	After (Framework)	Change
Duplicate record rate	~18% estimated	9.1% post-resolution	>50% reduction
Identity F1 score	~83.4% (det. only)	93.1% (hybrid)	+9.7 points
Fraud detection lift	1.00 (baseline)	2.31x	+131%
False-positive fraud flags	Baseline	-68% reduction	Significant improvement
Graph clustering C_eff	0.31 (det. only)	2.14 (hybrid)	+6.9x
GDPR compliance coverage	Partial (det. only)	Full (all records)	Complete
Processing throughput	~0.5M rec/s (legacy)	2.1M rec/s	+320%
Incremental update time	Full batch (hours)	<22 minutes	Near real-time

4. Conclusion

Identity resolution in the telecommunications industry is generally performed using a combination of deterministic, probabilistic and graph-based matching to meet the strict accuracy and scalability needs of production systems. Deterministic matching alone achieves high precision but insufficient recall because telecommunications data is inherently incomplete and subject to continuous identifier changes. Probabilistic matching improves recall at the cost of some precision reduction. Graph-based architectures expose relational patterns that neither pairwise comparison model can detect independently. The composite identity framework presented in this article combines all three modalities into a single scoring model, achieving 93.1 percent F1 accuracy, a 9.1 percent deduplication rate across 527 million records, graph clustering efficiency of 2.14, and a fraud detection lift of 2.31 due to graph-guided identity clustering. These results represent consistent improvement over any single-modality baseline across all evaluation metrics. Identity governance with role-based access control and federated authorization supports enterprise-level compliance with GDPR and CCPA regulations. The segment-specific calibration of matching parameters is performed through dynamic customer segmentation to address the heterogeneous data quality of prepaid, postpaid, and enterprise subscriber populations. Prepaid segments have the highest duplication rates and thus benefit the most from graph-based extensions. Future work includes graph transformer architectures to handle noisy matchings and real-time streaming graph updates using Apache Flink, enabling sub-minute identity updates, as well as federated machine learning approaches to allow for identity resolution across telecommunications operators while ensuring subscriber privacy under differential privacy requirements.

Acknowledgments

The author acknowledges the open-source research communities whose published work on entity resolution, distributed graph systems, and privacy-preserving record linkage informed the framework presented in this article. No external funding was received for this research.

Author Contributions

Suresh Tambe: Conceptualization, methodology, formal analysis, investigation, writing (original draft preparation), and writing (review and editing).

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Ivan P. Fellegi and Alan B. Sunter, "A Theory for Record Linkage", Journal of the American Statistical Association, 1969. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049

[2] Junwei Hu et al., "When GDD Meets GNN: A Knowledge-Driven Neural Connection for Effective Entity Resolution in Property Graphs", Information Systems, 2025. Available: https://www.sciencedirect.com/science/article/pii/S0306437925000365

[3] Mohammad Heydari et al., "Distributed Record Linkage in Healthcare Data with Apache Spark", arXiv preprint arXiv:2404.07939, 2024. Available: https://arxiv.org/abs/2404.07939

[4] Xinxin Hu et al., "GAT-COBO: Cost-Sensitive Graph Neural Network for Telecom Fraud Detection", IEEE Transactions on Big Data, 2024. Available: https://ieeexplore.ieee.org/abstract/document/10388428

[5] Neil G. Marchant et al., "Bayesian Graphical Entity Resolution Using Exchangeable Random Partition Priors", Journal of Survey Statistics and Methodology, 2023. Available: https://academic.oup.com/jssam/article/11/3/569/6969687

[6] Lasya Vyakaranam et al., "Simplifying Graph-Parallel Computation in Apache Spark with GraphX", 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2024. Available: https://ieeexplore.ieee.org/abstract/document/10800958

[7] Xinxin Hu et al., "Mining Mobile Network Fraudsters with Augmented Graph Neural Networks", Entropy, 2023. Available: https://www.mdpi.com/1099-4300/25/1/150

[8] Thanh Huan Vo et al., "Extending the Fellegi-Sunter Record Linkage Model for Mixed-Type Data with Application to the French National Health Data System", Computational Statistics and Data Analysis, 2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167947322002365

[9] Lavanya Elluri et al., "An Integrated Knowledge Graph to Automate GDPR and PCI DSS Compliance", 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018. Available: https://ieeexplore.ieee.org/abstract/document/8622236

[10] Alexandros Karakasidis and Georgia Koloniari, "Efficient Privacy Preserving Record Linkage at Scale Using Apache Spark", 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022. Available: https://ieeexplore.ieee.org/abstract/document/10020832

[11] Reynold S. Xin et al., "GraphX: A Resilient Distributed Graph System on Spark", First International Workshop on Graph Data Management Experiences and Systems (GRADES), ACM, 2013. Available: https://dl.acm.org/doi/abs/10.1145/2484425.2484427

[12] An Tong et al., "GDFGAT: Graph Attention Network Based on Feature Difference Weight Assignment for Telecom Fraud Detection", PLOS ONE, 2025. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0322004

[13] Olivier Binette and Rebecca C. Steorts, "(Almost) All of Entity Resolution", Science Advances, 2022. Available: https://www.science.org/doi/full/10.1126/sciadv.abi8021

[14] Junhang Wu et al., "Beyond the Individual: An Improved Telecom Fraud Detection Approach Based on Latent Synergy Graph Learning", Neural Networks, 2024. Available: https://www.sciencedirect.com/science/article/abs/pii/S0893608023005737

[15] Yuhang Zhang et al., "Scalable Entity Resolution Using Probabilistic Signatures on Parallel Databases", Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, 2018. Available: https://dl.acm.org/doi/abs/10.1145/3269206.3272016

[16] George Papadakis et al., "Three-Dimensional Entity Resolution with JedAI", Information Systems, 2020. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306437920300570

[17] Robert A. Barton et al., "Graph Neural Networks for Inconsistent Cluster Detection in Incremental Entity Resolution", arXiv preprint arXiv:2105.05957, 2021. Available: https://arxiv.org/abs/2105.05957

[18] Olalekan Hamed Olayinka, "Data Driven Customer Segmentation and Personalization Strategies in Modern Business Intelligence Frameworks", World Journal of Advanced Research

and Reviews, 2021. Available: https://doi.org/10.30574/wjarr.2021.12.3.0658

[19] Amir Ghasemian et al., "Evaluating Overfit and Underfit in Models of Network Community Structure", IEEE Transactions on Knowledge and Data Engineering, 2019. Available: https://ieeexplore.ieee.org/abstract/document/8692626

[20] George Papadakis et al., "A Critical Re-Evaluation of Record Linkage Benchmarks for Learning-Based Matching Algorithms", 2024 IEEE 40th International Conference on Data Engineering (ICDE), IEEE, 2024. Available: https://ieeexplore.ieee.org/abstract/document/10598022