Solyra: A Probabilistic Geospatial Deduplication and Triage Framework for Decision Support in High-Density Refugee Settlements

Author: Sohaib Khan
Date: May 2026

Keywords: Humanitarian Informatics; Crisis Mapping; Entity Resolution; Geospatial Analysis; Decision Support Systems; Humanitarian Logistics; Data Responsibility

---

Abstract

In a refugee settlement, reporting is often rough, duplicating, delayed and unclear, which leads to misuse or inefficient use of resources. This paper introduces Solyra which aims to solve this issue by probabilistic geospatial deduplication and triage framework.

Solyra integrates spatial, temporal, semantic, and source-reliability signals to estimate the likelihood that reports refer to the very same incident. The system produces a confidence-weighted "Living Map", which gives a clear and coordinated picture of issues to all the clusters involved.

This paper presents the system architecture, probabilistic model, triage logic, comparative positioning, evaluation design, and ethical framework.

---

1. Introduction

Humanitarian response environments, particularly refugee camps, are characterized by fragmented, overlapping, and delayed reporting. Multiple actors may report the same issue, while other critical issues remain underreported. This results in delays, misuse or overuse of the resources, unnecessary individualistic actions by multiple clusters.

Existing tools such as KoboToolbox and ArcGIS-based workflows are effective for data collection and visualization but rely heavily on manual deduplication, introducing delays and inconsistencies.

Contributions

This paper makes three contributions:

1. It formalizes humanitarian deduplication as a probabilistic entity resolution problem under spatiotemporal uncertainty.
2. It introduces a multi-signal clustering framework integrating spatial, temporal, semantic, and source-reliability features.
3. It proposes an uncertainty-aware triage model for humanitarian decision support.

---

2. Problem Formulation

Let reports be represented as:

$$R = \{r_1, r_2, ..., r_n\}$$

Each report is modeled as:

$$r_i = (x_i, y_i, t_i, a_i, s_i)$$

Where:

- (x_i, y_i): spatial coordinates
- (t_i): timestamp
- (a_i): event attributes and severity indicators
- (s_i): source metadata

The objective is to estimate:

$$P(r_i \sim r_j)$$

which represents the probability that two reports correspond to the same underlying incident.

This formulation aligns humanitarian deduplication with probabilistic entity resolution and record linkage problems studied in information systems research.

---

3. Deduplication Framework

The Solyra framework models report deduplication as a probabilistic entity resolution problem, where the objective is to estimate the likelihood that two reports correspond to the same underlying incident. Each report is represented as a structured tuple containing spatial coordinates, temporal information, semantic attributes, and source metadata. Rather than relying on deterministic matching rules, the framework evaluates similarity across multiple dimensions to capture the inherent uncertainty in humanitarian reporting environments.

Spatial similarity is derived from the geographic distance between reports, acknowledging that reporting inaccuracies and environmental constraints may introduce noise. Temporal similarity accounts for both reporting delays and the persistence of real-world events, recognizing that incidents may generate reports over extended time windows. Semantic similarity is computed based on event type, severity indicators, and descriptive features, enabling differentiation between co-located but distinct issues. Finally, source reliability is incorporated to weight reports according to their credibility, which may depend on reporter type, historical consistency, or cross-validation with other sources.

These dimensions are integrated into a probabilistic model that estimates the likelihood of equivalence between reports. The model combines spatial distance, temporal difference, semantic similarity, and source interaction into a unified scoring function, which is then transformed using a logistic function to produce a probability value.

$$P(r_i \sim r_j) = \sigma(\alpha d_{ij} + \beta \Delta t_{ij} + \gamma m_{ij} + \delta q_{ij})$$

Where:

- d_{ij}: spatial distance
- Δt_{ij}: temporal difference
- m_{ij}: semantic similarity
- q_{ij}: source reliability interaction
- σ: logistic function

Clustering is performed using a graph-based approach, where each report is treated as a node and edges are established between pairs of reports whose estimated probability exceeds a defined threshold:

$$P(r_i \sim r_j) > \theta$$

Event clusters are then identified as connected components within this graph. This approach avoids rigid assumptions about cluster shape and allows the model to adapt to varying spatial and reporting conditions.

To ensure scalability in high-density environments, the system employs approximate nearest-neighbor techniques to restrict pairwise comparisons to spatially relevant candidates. This reduces computational complexity from $O(n^2)$ to approximately $O(n \log n)$, making the framework feasible for large-scale deployments.

---

4. The Living Map and Triage Logic

Each event cluster forms an Event Node characterized by:

- Confidence score
- Source diversity
- Temporal evolution
- Uncertainty bounds

Solyra produces a dynamic “Living Map” where uncertainty is represented explicitly rather than hidden behind deterministic labels.

Triage States

Red

High-confidence actionable incidents supported by strong multi-source evidence.

Yellow

Uncertain or partially validated incidents requiring further verification.

Green

Resolved incidents retained for historical continuity and possible reactivation.

State transitions are reversible and evidence-driven, allowing the system to adapt as new reports arrive.

Figure 1. Solyra System Architecture

```
┌──────────────────────────────┐
|     Ingestion Layer       |
| SMS • Mobile • Web • APIs   |
└──────────────┬───────────────┘
          |
          ▼
┌──────────────────────────────┐
|      Preprocessing        |
| Cleaning • Parsing • Normal. |
```

```
└──────────────┬──────────────┘
       │
       ▼
┌─────────────────────────────┐
│    Feature Extraction       │
│ Spatial • Temporal • Semantic│
│ Source Reliability Signals  │
└──────────────┬──────────────┘
       │
       ▼
┌─────────────────────────────┐
│    Similarity Engine        │
│ Pairwise Probability Model  │
└──────────────┬──────────────┘
       │
       ▼
┌─────────────────────────────┐
│    Graph Construction       │
│ Nodes = Reports             │
│ Edges = Probabilistic Links │
└──────────────┬──────────────┘
       │
       ▼
┌─────────────────────────────┐
│      Clustering             │
│ Connected Components        │
│ Event Formation             │
└──────────────┬──────────────┘
```

```
          |
          ▼
┌────────────────────────────────┐
│       Event Nodes        │
│ Confidence • History • State │
└───────────────┬────────────────┘
          |
          ▼
┌────────────────────────────────┐
│      Visualization       │
│ Uncertainty-Aware Living Map │
└────────────────────────────────┘
```

Caption:
Figure 1. End-to-end Solyra architecture showing ingestion, probabilistic deduplication, graph-based clustering, and uncertainty-aware visualization.

Figure 2. Probabilistic Deduplication Pipeline

```
┌──────────────┐
│  Reports  │
└──────┬───────┘
     |
     ▼
┌─────────────────────┐
│ Feature Vectors  │
│ Spatial Coordinates│
│ Timestamps       │
│ Semantic Features │
│ Source Metadata   │
└──────┬──────────────┘
```

```
      |
      ▼
┌──────────────────┐
| Pairwise Similarity|
|  Probability Model |
|                  |
| P(ri ~ rj)       |
└───────┬──────────┘
      |
      ▼
┌──────────────────┐
| Graph Construction |
| Nodes = Reports    |
| Edges = Similarity |
└───────┬──────────┘
      |
      ▼
┌──────────────────┐
|   Clustering       |
| Connected Events   |
└───────┬──────────┘
      |
      ▼
┌──────────────────┐
|  Triage Engine     |
| Red • Yellow • Green|
└──────────────────┘
```

Caption:
Figure 2. Deduplication pipeline illustrating probabilistic linkage, graph construction, clustering, and triage generation.

Figure 3. Event Lifecycle and State Transitions

```
┌─────────────────┐
│     Yellow      │
│    Uncertain    │
│ Partial Evidence │
└────────┬─────────┘
         │
Additional │ Validation
 Evidence  ▼
┌─────────────────┐
│      Red        │
│ High Confidence  │
│ Operational Alert│
└────────┬─────────┘
         │
  Resolution│
         ▼
┌─────────────────┐
│     Green       │
│    Resolved     │
│ Historical State │
└────────┬─────────┘
         │
  New Data │ Reactivation
```

```
        ▼
  ┌───────────────┐
  │    Yellow     │
  │ Reopened Cluster │
  └───────────────┘
```

Caption:
Figure 3. Event lifecycle showing reversible confidence-driven transitions between uncertain, actionable, and resolved states.

Figure 4. Multi-Signal Similarity Model

```
┌──────────────────────┐
│    Spatial Distance     │
└──────────┬───────────┘
           │
┌──────────────────────┐
│    Temporal Distance    │
└──────────┬───────────┘
           │
┌──────────────────────┐
│    Semantic Similarity  │
└──────────┬───────────┘
           │
┌──────────────────────┐
│    Source Reliability   │
└──────────┬───────────┘
           │
           ▼
┌──────────────────────┐
```

```
| Logistic Probability Model |
|                   |
| P(ri ~ rj)             |
└──────────────┬────────────────┘
        |
        ▼
┌────────────────────────────────┐
| Event Link / No Link      |
└────────────────────────────────┘
```

Caption:
Figure 4. Multi-dimensional similarity model integrating spatial, temporal, semantic, and source-reliability signals.

5. System Architecture

The system architecture consists of four primary layers.

The ingestion layer accepts reports from multiple channels, including SMS, mobile devices, web interfaces, and API integrations. This flexibility is important in low-connectivity humanitarian environments where communication infrastructure may be unreliable.

The preprocessing layer performs data cleaning, normalization, and feature extraction. This includes parsing location data, timestamps, and semantic attributes from incoming reports.

The deduplication engine performs probabilistic linkage and graph-based clustering. Reports are continuously evaluated and merged into evolving event nodes based on similarity estimates and confidence thresholds.

The analysis layer supports trend analysis, recurrence detection, and operational simulation, while the visualization layer provides a dynamic uncertainty-aware map that supports situational awareness for field operators and coordinators.

The system is designed for low-bandwidth deployment scenarios and supports asynchronous synchronization where persistent connectivity is unavailable.

---

6. Comparative Positioning

Feature	Legacy GIS Workflows	Solyra
Deduplication	Manual	Probabilistic
Triage	Subjective	Structured
Uncertainty	Implicit	Explicit
Decision Speed	Delayed	Faster
Adaptability	Static	Dynamic

Solyra is not intended to replace existing humanitarian GIS systems. Instead, it functions as a decision-support augmentation layer designed to improve operational clarity under uncertainty.

---

7. Real-World Applicability

The framework is designed for high-density refugee camps, informal settlements, and disaster-response environments where repetetive reporting is common.

Several operational constraints influence deployment.

First, co-located incidents may occur within small geographic areas, requiring semantic differentiation between otherwise spatially similar reports. Second, reporting behavior may be socially amplified, leading to overrepresentation of highly visible incidents. Third, temporal drift and recurrence patterns complicate event persistence modeling. Finally, infrastructure limitations may constrain connectivity and synchronization reliability.

Solyra addresses these constraints through adaptive clustering, source weighting, temporal decay modeling, and offline-first synchronization strategies.

---

8. Cost-Effectiveness

Solyra hypothesizes that efficiency gains emerge primarily through reductions in redundant verification activities and improved prioritization of field operations.

Potential operational benefits include:

- Reduced duplicate site visits
- Faster incident prioritization
- Lower coordination overhead
- Improved allocation of limited humanitarian resources

However, these gains are context-dependent and vary according to the baseline level of report duplication and organizational workflow maturity.

The framework therefore defines efficiency improvements as conditional hypotheses requiring empirical validation rather than universal guarantees.

---

9. Evaluation Framework

The evaluation of Solyra is designed to assess both technical deduplication performance and operational impact. Given the difficulty of obtaining fully labeled humanitarian datasets, the framework combines synthetic simulation with partial ground-truth validation.

Reports may be drawn from historical humanitarian datasets or generated synthetically to reflect realistic spatial clustering, reporting delays, and duplication patterns. A subset of reports is manually annotated to distinguish between true duplicates, co-located events, and independent incidents.

Performance is evaluated using standard entity-resolution metrics including pairwise precision, recall, cluster purity, and fragmentation. Particular attention is given to the false merge rate, since incorrect aggregation of distinct incidents may suppress critical operational signals.

An operational cost model is also introduced:

$$C = c_v V + c_e E$$

Where:

- (V): validation site visits
- (E): error corrections
- (c_v), (c_e): associated operational costs

Simulation-based evaluation is intended to explore the potential operational effects of probabilistic deduplication under varying reporting conditions rather than to establish definitive performance benchmarks prior to field deployment.

---

10. Ethical and Governance Framework

The framework follows a privacy-by-design approach and does not require personally identifiable information for core functionality. Data collection is minimized to operationally relevant fields necessary for probabilistic linkage and triage.

Transparency is emphasized throughout the system design. Confidence scores and uncertainty representations are exposed explicitly to users rather than concealed behind deterministic classifications. This reduces the risk of false certainty and supports accountable decision-making.

The system is intended as a decision-support tool rather than an authoritative autonomous system. Human operators retain final responsibility for interpretation and intervention prioritization.

The framework aligns conceptually with:

- UN OCHA Data Responsibility Guidelines
- IASC Operational Guidance
- GDPR principles of proportionality and minimization

Deployment should include local ethical review, contextual calibration, and community engagement processes appropriate to humanitarian environments.

---

11. Limitations

Solyra relies on assumptions regarding data availability, reporting diversity, and feature quality. In sparse or adversarial reporting environments, probabilistic linkage performance may degrade significantly.

The framework also depends on local calibration to account for differing reporting behaviors, infrastructure layouts, and organizational workflows. Additionally, simulation-based evaluation cannot fully capture the social and operational complexities of real humanitarian deployments.

Finally, while probabilistic deduplication may improve efficiency, aggressive merging strategies introduce epistemic risks that require careful management.

---

12. Discussion

The implementation of probabilistic deduplication in humanitarian contexts introduces a fundamental trade-off between operational efficiency and epistemic risk. While consolidating redundant reports can reduce unnecessary field visits and improve response prioritization, incorrect merges may obscure distinct incidents and lead to under-response in critical areas. This tension highlights the importance of balancing automation with transparency.

Solyra addresses this challenge by explicitly representing uncertainty rather than attempting to eliminate it. By assigning confidence scores to event nodes and maintaining reversible state transitions, the system allows decision-makers to interpret outputs in context rather than treating them as definitive truths. This design aligns with the principle that decision-support systems in humanitarian settings should augment, rather than replace, human judgment.

Another key consideration is the variability of reporting behavior across different environments. Factors such as population density, communication access, and organizational coordination can significantly influence the distribution and quality of incoming reports. As a result, the effectiveness of probabilistic deduplication depends on careful calibration to local conditions. A model that performs well in one context may require adjustment in another to maintain reliability.

The framework also raises important questions about trust and adoption. For field operators to rely on probabilistic outputs, the system must provide interpretable signals and avoid presenting overly precise conclusions. By emphasizing transparency and exposing uncertainty, Solyra seeks to build confidence in its recommendations while preserving the flexibility needed for complex decision-making.

Overall, the discussion underscores that the value of Solyra lies not only in its ability to reduce redundancy, but in its capacity to structure uncertainty in a way that supports informed, context-sensitive humanitarian action.

---

13. Conclusion

At present, Solyra is a conceptual and simulation-oriented framework and has not yet undergone operational deployment in live humanitarian settings. It proposes a shift from deterministic mapping toward probabilistic, uncertainty-aware decision support in humanitarian operations.

Its primary contribution lies in structuring redundant information while making uncertainty explicit rather than hidden. By combining probabilistic linkage, adaptive clustering, and confidence-aware triage, the framework supports more informed operational decision-making in environments characterized by incomplete and overlapping information.

Future work should focus on real-world pilot deployments, empirical calibration, and integration with existing humanitarian information systems.

References

Fellegi, I. P., & Sunter, A. B. (1969). *A Theory for Record Linkage*. Journal of the American Statistical Association, 64(328), 1183–1210.
https://www.jstor.org/stable/2286061

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer. https://link.springer.com/book/10.1007/978-3-642-31164-2

Winkler, W. E. (2006). *Overview of Record Linkage and Current Research Directions*. U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2006/adrm/rrs2006-02.pdf

Goodchild, M. F. (2007). *Citizens as Sensors: The World of Volunteered Geography*. GeoJournal, 69(4), 211–221. https://link.springer.com/article/10.1007/s10708-007-9111-y

European Union. (2016). *General Data Protection Regulation (GDPR)*. https://gdpr-info.eu/

ESRI. *ArcGIS Documentation*. https://www.esri.com/en-us/arcgis/products/index

KoboToolbox. *Documentation*. https://www.kobotoolbox.org