

From Hong Kong Counter-Trafficking to Billions of Comparisons a Day: Practical Entity Resolution When Records Are Noisy or Adversarial

Taylor Amarel · Senior Data Scientist (IVANS) · Creator of [bq-entity-resolution](#) · Portal: <https://papers.baltor.ai>

Overview: Whether you are matching millions of commercial insurance policies across carrier feeds at IVANS or linking shell recruitment agencies across Hong Kong corporate registries at Migrasia, the core engineering problem is the same: **you cannot compare every record to every other record ($O(N^2)$), and the records you care about most are the ones where names, addresses, and identifiers do not match cleanly.** This paper distills practical lessons from building [bq-entity-resolution](#) and investigative entity graphs.

1. TWO WORLDS OF ENTITY RESOLUTION: ACCIDENTAL NOISE VS. DELIBERATE CAMOUFLAGE

Dimension	Enterprise Commercial Data (IVANS Scale)	Counter-Trafficking & Illicit Network Investigations (Migrasia / DueCare)
Scale & Throughput	Billions of candidate pairs per day across warehouse tables	Tens of thousands of agencies, lenders, directors, phone numbers, and job orders
Why Records Diverge	Typos, DBA trade names, abbreviations, parent/subsidiary splits, stale addresses	Deliberate phoenixing: re-registering closed agencies under a relative's name while keeping the same phone, fax, or suite
Primary Blocking Signals	Normalized name tokens, phonetic keys, ZIP5/geohash, FEIN/domain prefixes	Shared mobile/WhatsApp numbers, building room coordinates, license transfer chains, contract boilerplate fingerprints
Cost of a False Positive	Merged distinct businesses; corrupted policy routing	Unfairly implicating a legitimate agency that happens to rent an office in the same commercial building

2. THE 5-STAGE PIPELINE IN `BQ-ENTITY-RESOLUTION`

- Stage 1 (Deterministic Hygiene & Legal-Suffix Peeling):** Strip corporate noise (`LLC` , `INC` , `LIMITED` , `EMPLOYMENT AGENCY`), standardize street/suite tokens, and canonicalize phone/tax IDs before touching any similarity function.
- Stage 2 (Multi-Pass Overlapping Blocking Canopies):** A single blocking key misses records when a typo lands in the first three characters. We union 4–6 narrow blocking passes (token-sort prefix, phonetic soundex/metaphone, address geohash + suite, and shared contact handle) so recall stays above 99% while pruning >99.9% of $O(N^2)$ pairs.
- Stage 3 (Composite Feature Scoring):** Combine Jaro-Winkler on short names, token-IDF cosine similarity on long business names, exact/partial phone and suite agreement, and domain-specific rarity weights.
- Stage 4 (Graph Transitive Closure & Bridge-Node Pruning):** Connected components ($A \sim B, B \sim C \Rightarrow A \sim C$) can explode into giant hairballs when a shared registered-agent address links 500 unrelated companies. We detect high-degree hub attributes (registered agents, shared secretarial suites) and down-weight them before clustering.
- Stage 5 (Calibrated Abstention & Audit Receipts):** Every resolved cluster carries the exact feature vector and edge weights that formed it, routing borderline pairs to human or LLM competing-hypothesis review.