

Auditable Corridor Shortlisting under Partial Observability: A Systems Approach to Ranking Human Trafficking Origin–Destination Links

Michael A Magid

mmagid1@binghamton.edu

Binghamton University

Melissa Zeynep Ertem

Binghamton University

David Cingranelli

Binghamton University

Todd Landman

University of Nottingham

Research Article

Keywords: complex networks, link prediction, human trafficking, partial observability, graph neural networks, algorithmic auditing

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11244867/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Auditable Corridor Shortlisting under Partial Observability: A Systems Approach to Ranking Human Trafficking Origin–Destination Links

Michael A. Magid^{1,3*}, Melissa Zeynep Ertem¹, David Cingranelli²,
Todd Landman³

¹Department of Systems Science and Industrial Engineering,
Binghamton University, Binghamton, 13902, NY, USA.

²Department of Political Science, Binghamton University, Binghamton,
13902, NY, USA.

³School of Politics and International Relations, University of
Nottingham, Nottingham, NG7 2RD, UK.

*Corresponding author(s). E-mail(s): mmagid1@binghamton.edu;

Abstract

International human trafficking is only partially observed at the corridor level: a small number of origin–destination (OD) pairs appear in public victim-case compilations, while the space of possible corridors is vast. We treat corridor prioritization as a systems problem and present an auditable ranking loop that maps sparse observed ODs, country-level covariates, and an explicit graph representation into a shortlist of corridors for subsequent review. The system couples three inference pathways (tabular attributes, graph neural message passing, and a temporal world model) under a frozen holdout and hard-matched negatives. Assurance is provided by a falsification battery (label destruction, structural nulls, graph rewiring, world-model ablations), calibration diagnostics, multi-seed stability checks, and prospective recovery of Counter-Trafficking Data Collaborative (CTDC) corridors first observed after year cuts. Across three seeds, the core battery passes on average 6.33 of 8 controls; graph-rewire tests pass on one of three seeds and are reported as unstable, and world-model interventions pass only on training-loss gaps. The tabular pathway is the strongest ranker (edge precision–recall area under the curve, PR-AUC, 0.59 ± 0.17 versus 0.48 ± 0.10)

for the graph neural network), but test splits contain roughly a dozen positives and seed variance is large. In prospective recovery, tabular recall@20 averages 0.50 over seeds and cuts, against chance PR-AUC of 0.25–0.37. The deliverable is not a certified map of trafficking routes but a reproducible shortlist with an attached verification record, designed to interface with external review.

Keywords: complex networks, link prediction, human trafficking, partial observability, graph neural networks, algorithmic auditing

1 Introduction

Human trafficking systems couple recruitment, transit, and exploitation across borders, yet measurement at the bilateral (corridor) scale remains sparse. Public compilations such as the Counter-Trafficking Data Collaborative (CTDC) record only a limited set of citizenship–exploitation country pairs, while analysts must still decide which corridors merit further inspection. Under partial observability, the decision-relevant output is an ordered set of corridors: a shortlist. The corridor network is a complex system: heterogeneous, relational, and observed only through institutional detection. Network science, and link prediction in particular, is a natural lens.

We develop this shortlisting task as a systems design. We specify (i) a state representation built from observed CTDC OD labels, Human Trafficking Indicators (HTI)-derived country context, and aligned covariates; (ii) coupled ranking mechanisms that convert state into corridor scores; and (iii) an assurance layer that intervenes on labels, graph structure, and temporal heads to test whether the shortlist depends on the structures claimed. Graph neural networks (GNNs) and a temporal world model appear as relational and dynamic pathways inside the loop.

A shorter version of this work, centered on temporal graph networks, is in press in the proceedings of the International Symposium on Complex Systems (ISCS 2026)

([Magid et al. 2026](#)). The present article reframes the problem as auditable shortlisting and adds hard-matched negatives, a comparison of three ranking pathways, the assurance battery, prospective evaluation, and the related work and ethics discussion.

2 Related work

Our setting sits at the intersection of four literatures: the measurement of trafficking under selective detection, quantitative and computational work on trafficking and migration flows, link prediction on partially observed networks, and verification practices for predictive systems.

2.1 Measuring trafficking under selective observation

Trafficking records are products of detection, identification, and reporting as much as of the underlying phenomenon. [Feingold \(2010\)](#) argues that circulating trafficking figures are socially constructed through institutional data production, [Tyldum \(2010\)](#) reviews the methodological limits of empirical trafficking research, and [Weitzer \(2014\)](#) questions macro-level estimates that cannot be substantiated and argues for more careful, smaller-scale evidence. [Shelley \(2014\)](#) argues that weak borders, corruption, and unmet labor demand make smuggling and trafficking into Europe profitable and low-risk, so which routes are visible depends on institutional context as well as on where movement occurs. A statistical response to partial observation is multiple-systems estimation, which combines overlapping lists to estimate unobserved totals ([Silverman 2020](#)); it estimates how many victims are missing, not which origin–destination corridors are. The data sources we use, the CTDC victim-case records ([Counter-Trafficking Data Collaborative 2017](#)) and the HTI country-year dataset ([Frank 2022](#)), inherit these selection effects. We therefore treat observed corridors as a biased sample of true corridors and avoid interpreting absence from the record as absence of a route.

139 2.2 Quantitative and computational approaches to trafficking 140 and flows 141

142
143 [Dimas et al. \(2022\)](#) classify 142 studies applying operations research and analytics to
144 trafficking. They report a strong skew toward sex trafficking (47.9% of studies ver-
145 sus 12.0% for labor trafficking), a concentration of sex trafficking studies in developed
146 countries that they attribute largely to the availability of online escort data, and
147 network-based methods in 34.5% of studies, used to detect and characterize observed
148 trafficking networks. Recent machine learning work scores risk from observable sig-
149 nals; for example, [Ramchandani et al. \(2025\)](#) analyze commercial sex advertisements
150 to identify risk in sex-industry supply chains. Such approaches score or characterize
151 entities and networks that appear in the data (advertisements, businesses, countries,
152 or observed links), whereas we rank dyads that are not in the record. In migration
153 research, gravity models are the standard tool for bilateral flows ([Beine et al. 2016](#)), but
154 [Beyer et al. \(2022\)](#) show they explain spatial variation across country pairs far better
155 than temporal variation within a pair, and cannot predict dynamics. Geographic dis-
156 tance is therefore a strong but incomplete prior for corridor structure. Network studies
157 of other illicit trades show similar geographic structure. [Arroyave et al. \(2020\)](#) find
158 that Colombian reptile trafficking routes track departmental adjacency and regional
159 proximity, and [Meneghini et al. \(2020\)](#) reconstruct the global illicit cigarette network
160 across 158 countries by combining seizure data with consumption and geographic infor-
161 mation. We use distance and degree-based scores as explicit null models rather than
162 as the model of interest.

177 2.3 Link prediction with missing edges and relational learning 178

179 Corridor ranking is a link-prediction problem on a network whose edges are partially
180 observed. Classical similarity and topology-based methods are surveyed by [Liben-
181 Nowell and Kleinberg \(2007\)](#) and [Lü and Zhou \(2011\)](#), and [Clauset et al. \(2008\)](#) show
182
183
184

that hierarchical network structure supports the prediction of missing links. Graph neural networks learn node representations by message passing (Kipf and Welling 2017; Veličković et al. 2018), and Zhang and Chen (2018) adapt them to link prediction using enclosing subgraphs. Temporal extensions include temporal graph networks (Rossi et al. 2020) and the families surveyed by Kazemi et al. (2020), and multilayer formalisms (Kivelä et al. 2014) motivate treating exploitation types or time slices as coupled layers. In our case, unobserved dyads are unlabeled rather than negative, which places the problem in the positive-unlabeled learning setting (Bekker and Davis 2020), where sampled non-edges are a convenience rather than ground truth. Evaluation choices then matter a great deal: Shchur et al. (2018) show that conclusions about graph neural networks depend on data splits and tuning, and Li et al. (2023) argue that common link-prediction evaluation settings are unrealistic and propose benchmarks with harder negative candidates. These observations motivate our use of a frozen holdout with a leakage audit and of negatives matched to positives on distance and destination severity, so that a model cannot succeed simply by separating nearby, high-severity pairs from random distant ones.

2.4 Verification, null models, and responsible use

Model scores are only useful if they are interpretable as such. We report calibration with expected calibration error and Brier score following Guo et al. (2017), who document that modern neural networks can be miscalibrated. To test whether a relational model depends on the claimed structure, degree-preserving randomization is a standard null (Fosdick et al. 2018), and we apply configuration-model rewiring to the message-passing graph. More broadly, we adopt the logic of internal algorithmic auditing (Raji et al. 2020): claims about what drives the output are tied to explicit interventions whose results, including failures, are retained in the record. Finally, predictions built from crime-related administrative data can inherit and amplify reporting

bias (Lum and Isaac 2016), so corridor rankings should be read with that bias in mind (Section 8).

2.5 Positioning

Prior work largely estimates how much trafficking is unobserved, scores or characterizes observable entities and networks, or predicts flows with geography-driven models. To our knowledge, few treat the output as a shortlist of unobserved corridors paired with an intervention-based verification record. This paper contributes that system view: a corridor shortlist produced under explicit data contracts (observed labels only, matched negatives, a non-label graph scaffold), an assurance battery whose failures are reported rather than filtered, and an evaluation that includes prospective recovery of corridors first recorded after a cut year.

3 System problem

Let $\mathcal{C}$ be a finite set of countries and $\mathcal{E}^* \subset \mathcal{C} \times \mathcal{C}$ the unknown set of trafficking-relevant directed corridors. Observation yields only a sparse labeled set $\mathcal{E}_{\text{obs}} \subset \mathcal{E}^*$ (CTDC ODs after harmonization), together with node-level and dyadic covariates. The feasible action is selection of a size- K shortlist S_K from a candidate pool $\mathcal{U} \supseteq \mathcal{E}_{\text{obs}}$.

We design a ranking system f that produces scores $s(u)$ for $u \in \mathcal{U}$ and sets

$$S_K = \text{TopK}_{u \in \mathcal{U}} s(u). \quad (1)$$

Ranking supervision uses observed CTDC ODs only; dense geographic constructions are not treated as trafficking edge labels. Unobserved candidates are matched on distance and severity bins so ranking cannot succeed merely by separating nearby high-severity pairs from random far pairs, and sampled negatives are unlabeled, not confirmed absences. Geographic k -nearest-neighbor (k NN) links may support message

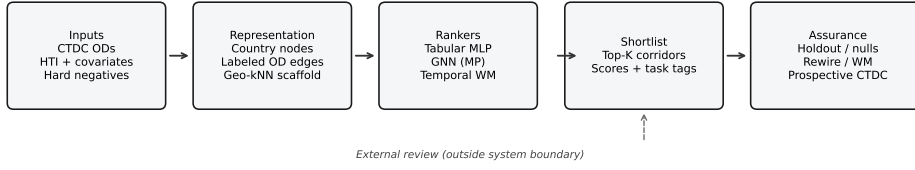


Fig. 1 Auditable corridor shortlisting system: inputs and representation feed coupled rankers, the shortlist is the decision interface, and assurance tests intervene before external review.

passing but are not counted as observed corridors. Claims about what informs S_K are tied to intervention tests (shuffle, null, rewire, ablation, or prospective cut).

Figure 1 summarizes the closed loop from inputs through shortlist to assurance, with external review outside the system boundary.

4 State representation

Nodes.

Countries in an HTI-aligned universe carry coordinates and context features (trafficking presence and roles, GDP and related covariates). The Human Trafficking Indicators dataset (Frank 2022) provides country-year data; in the reported run the 2017 cross-section supplies node context. Synthetic GDP fills are flagged; about 12% of temporal feature rows use any synthetic GDP flag.

Labeled edges.

Positives are unique CTDC origin–destination pairs (citizenship $\rightarrow$ country of exploitation) mapped into the country universe ($n_{\text{pos}} = 74$). The harmonized CTDC records used here span registration years 2002–2018.

Table 1 Data contract for the reporting protocol.

Quantity	Value
Countries in universe	179
Primary observed OD source	CTDC (harmonized), registration years 2002–2018
Labeled OD positives	74
Hard-matched negatives	148 (ratio 2)
Matching	1,000 km distance bins; 4 destination-severity bins
HTI role	Node/context features (not positive edge census)
Geo- k NN	Message-passing scaffold only
Temporal feature rows	1,640
Rows with any synthetic GDP flag	$\approx 12.2\%$
Holdout	Frozen train/val/test pairs + leakage audit
Leakage failures (seeds 0–2)	0

Negatives.

Hard-matched unobserved ODs are sampled at ratio 2 relative to positives ($n_{\text{neg}} = 148$), matched on distance (1,000 km bins) and destination-severity (four bins) (Table 1). Under this ratio the chance level of the area under the precision–recall curve (PR-AUC) on edge ranking is 1/3.

Scaffold.

For relational models, a geo- k NN graph may be unioned with observed flow edges for message passing. Scaffold edges are not ranking labels (Figure 2).

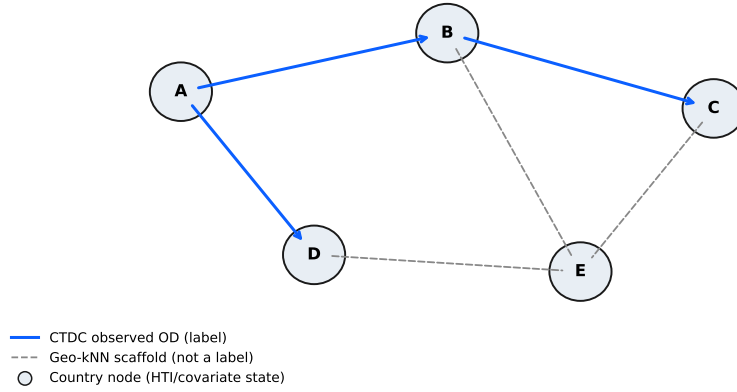
Time.

Pairs are embedded in temporal feature rows under a frozen train/validation/test holdout. Nested tasks (held-out edge ranking, new-corridor, hidden-current, future-horizon) re-slice the same scores; they do not redefine the shortlist product.

5 Ranking mechanisms

Three pathways map state to corridor scores under the same holdout and candidate contract.

Graph representation for corridor ranking



Hard-matched unobserved ODs enter as negatives in the ranking pool (not drawn).

Fig. 2 Graph representation: CTDC ODs as labeled directed edges, geo- k NN as non-label scaffold, nodes carrying HTI and covariate state. Hard negatives populate the ranking pool without being drawn as observed structure.

Attribute pathway (tabular).

A multilayer perceptron (MLP) scores OD pairs from node and dyadic covariates. Under the reporting protocol, geographic and degree shortcuts are dropped from the primary tabular feature set so the baseline is not a disguised distance model.

Relational pathway (GNN).

A graph neural network with message passing over observed flow edges (union geo- k NN) produces node states; an edge decoder scores candidates. In the structure-sensitive setting, geographic edge features and distance-based priors are zeroed in the decoder so ranking cannot bypass topology.

Dynamic pathway (temporal world model).

A country-year gated recurrent unit (GRU) maintains latent trajectories; transition and burden heads regularize dynamics; edge scores may use transitioned latents and burden features when enabled.

415 All pathways emit scores on the same test candidates; the system shortlist is TopK
416 of a chosen pathway. We report pathways separately to expose which channel carries
417 signal.
418
419

420 421 **6 Assurance layer** 422

423 Assurance asks whether S_K is using the intended system structure.
424
425

426 *Integrity.* 427

428 Frozen holdout artifacts and leakage audits precede fitting. Coordinate coverage, edge-
429 source counts, negative-source counts, and synthetic covariate fractions are recorded.
430
431

432 *Label and null interventions.* 433

434 Train-label permutation must collapse ranking if declared positives matter. Degree-
435 product and geography null scores must underperform fitted scores if the shortlist is
436 not reducible to preferential attachment or distance.
437
438

439 *Relational interventions.* 440

441 Configuration-model rewiring (Fosdick et al. 2018) of the message-passing edge set,
442 with a structure-sensitive decoder and path/recovery metrics under a *fixed* known
443 train graph, tests topology dependence. Multi-seed evaluation is required.
444
445

446 *Dynamic interventions.* 447

448 Frozen transition and shuffled burden targets test whether dynamics and burden heads
449 shape scores when on the ranking path. Latent time-shuffle collapse ratios probe
450 temporal structure in learned states.
451
452
453
454
455
456
457
458
459
460

Prospective CTDC cuts.

For cut year t , OD pairs first seen at or before t are known; pairs first seen after t are prospective targets. Rankers are assessed on recovery among test candidates (PR-AUC and recall@ K). This is public-data coherence, not operational confirmation.

Calibration and seeds.

Expected calibration error (ECE) and Brier score (Guo et al. 2017) accompany ranking metrics. Core assurance controls are repeated on seeds $\{0, 1, 2\}$. Two further controls (year-embedding shuffle and train-year permutation) are run and disclosed but excluded from the core eight, since they probe the year features rather than the shortlist’s main structure.

7 Empirical setting and results

Setting.

As in Table 1. Rankers: tabular MLP; GNN (flow $\cup$ geo- k NN, structure-sensitive); temporal world model. Seeds $\{0, 1, 2\}$. Each test split contains roughly a dozen positive corridors (12–13 across seeds), so single-seed metrics are noisy; we therefore report all seeds and avoid fine rankings between pathways.

Ranking metrics.

Table 2 reports multi-seed means and standard deviations. Chance PR-AUC is $1/3$, and chance area under the receiver operating characteristic curve (AUC-ROC) is 0.5. The tabular pathway is the strongest ranker, but its seed-to-seed spread is large (edge PR-AUC from 0.48 to 0.78). The GNN is moderately above chance on PR-AUC and close to chance on AUC-ROC, and the world model is at chance on AUC-ROC. Absolute PR-AUC is moderate under hard negatives and sparse labels. Tabular ECE of 0.25 indicates scores should be read as an ordering rather than as probabilities. Brier

Table 2 Primary ranking metrics under the reporting protocol (mean $\pm$ std over seeds $\{0, 1, 2\}$).
 Chance: PR-AUC = 1/3, AUC-ROC = 0.5.

Pathway	Edge PR-AUC	Edge AUC-ROC	ECE	Brier
Tabular MLP	0.589 ± 0.168	0.695 ± 0.144	0.249 ± 0.068	0.244 ± 0.072
GNN (structure-sensitive)	0.482 ± 0.104	0.557 ± 0.077	0.366 ± 0.015	0.357 ± 0.022
Temporal world model	0.427 ± 0.014	0.494 ± 0.047	0.039 ± 0.024	0.242 ± 0.006

scores are similar for the tabular pathway and the world model (0.24) and higher for the GNN (0.36).

Assurance outcomes.

Mean core pass rate is 6.33/8 (Table 3; Figure 3). Degree-product and geography nulls pass on all three seeds, so the fitted tabular scores are not reducible to those baselines in this protocol. The label-shuffle control passes on two of three seeds; on seed 1 the shuffled-label null scored higher than the fitted model (PR-AUC 0.67 versus 0.48), consistent with that seed’s near-chance tabular performance. Graph rewiring is not stable: both rewire PR-AUC and rewire structure pass on only one of three seeds. We therefore do not claim robust topology dependence of the shortlist at current OD scale; relational results are boundary conditions, not settled facts.

The two world-model controls pass on all seeds, but the passes rest on training-loss gaps. The change in edge PR-AUC under frozen transition or shuffled burden targets is at most 0.013, below the 0.02 threshold, and negative on seed 2. We do not treat these as confirmation that the dynamic heads shape scores. Moreover, the world-model future-route metric is inert in this protocol: on rows with a defined future label, the future label coincides with the current label, world-model scores are rank-identical to negative log distance, and the full and frozen-transition models are rank-identical, so the future-route delta is exactly zero. We therefore report no future-route world-model effects. The transition-collapse control (time-shuffled latents increase reconstruction

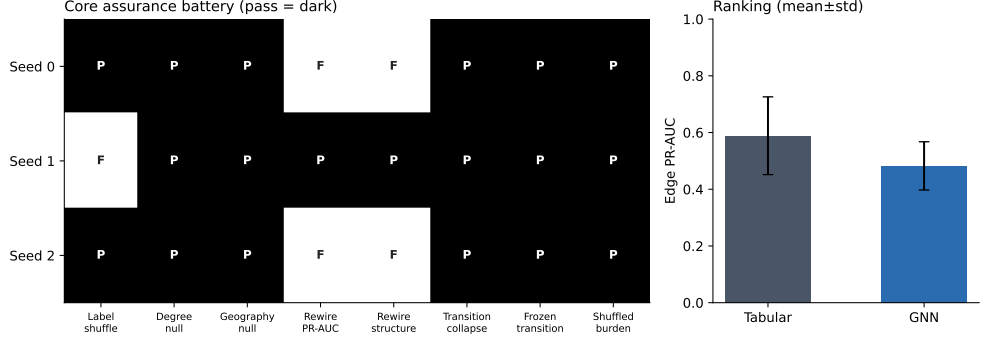


Fig. 3 Left: core assurance battery across seeds (P, pass; F, fail). Right: tabular versus GNN edge PR-AUC (mean±std). Rewire controls fail on two of three seeds.

Table 3 Assurance pass rates across seeds $\{0, 1, 2\}$ ($\delta_{\min} = 0.02$ where applicable). The first eight rows are the core battery. ^aWorld-model passes rest on training-loss gaps (see text).

Control	Passes / 3	Pass rate
Train label shuffle	2	0.67
Degree-product null	3	1.00
Geography null	3	1.00
Graph rewire (PR-AUC)	1	0.33
Graph rewire (structure)	1	0.33
Transition collapse	3	1.00
World-model frozen transition ^a	3	1.00
World-model shuffled burden ^a	3	1.00
Mean core pass count	6.33/8	
<i>Supplementary controls (not in core battery)</i>		
Year-embedding shuffle	1	0.33
Train-year permutation	1	0.33

error) passes on all seeds and shows that the learned latents carry temporal structure, which is a weaker statement than a ranking dependence.

Finally, of the two supplementary controls, the year-embedding shuffle passes on one of three seeds and the train-year permutation on one of three. Year features appear largely inert for ranking in this protocol, which tempers any reading of the pipeline as strongly temporal.

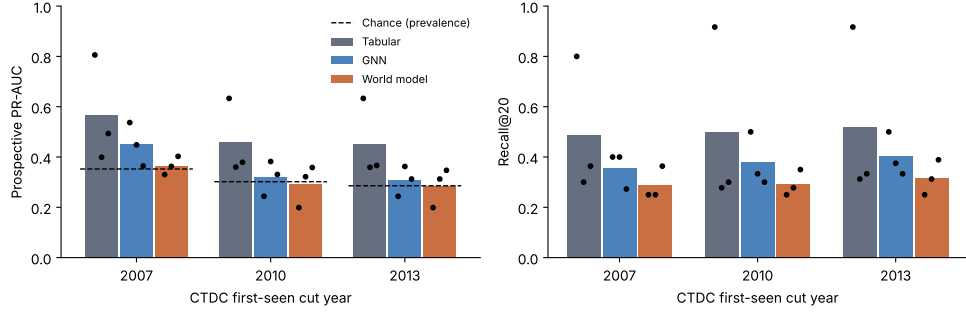


Fig. 4 Prospective CTDC recovery by cut year. Bars show means over seeds 0–2, dots show individual seeds, and dashed lines mark chance PR-AUC.

Prospective recovery.

Figure 4 and Table 4 summarize prospective CTDC cuts at 2007, 2010, and 2013 for all three seeds. Candidate pools are small (48–60 pairs) and contain 12–22 of the 37–50 prospective ODs, so only part of the prospective set is in the candidate frame and chance PR-AUC (prevalence) ranges from 0.25 to 0.37. Averaged over seeds and cuts, the tabular pathway leads (PR-AUC 0.49, recall@20 0.50), followed by the GNN (0.36, 0.38) and the world model (0.32, 0.30). The advantage is driven by seed 0; on seeds 1 and 2 the tabular pathway is only modestly above chance and within noise of the GNN, and on seed 1 the GNN has slightly higher PR-AUC. Attribute structure therefore carries somewhat more prospective hit rate than relational dynamics at this OD scale, but the margin is not robust across seeds.

On seed 0 the 2010 and 2013 cuts yield identical metrics. This is a holdout-coverage effect, not a pipeline fault: the five CTDC ODs that move from prospective to known between 2010 and 2013 are not in the seed-0 holdout test frame, so the candidate mask, labels, and scores coincide. Evaluating cuts on a cut-specific OD frame rather than the random test slice would remove this artifact and is left to future work.

Table 4 Prospective CTDC recovery averaged over cuts {2007, 2010, 2013} and seeds (per-seed means in parentheses: seeds 0, 1, 2). Chance PR-AUC (prevalence) is 0.25–0.37 (mean 0.31).

Pathway	Mean PR-AUC	Mean recall@20	Mean ECE
Tabular MLP	0.492 (0.69, 0.37, 0.41)	0.502 (0.88, 0.30, 0.33)	0.223
GNN	0.358 (0.34, 0.40, 0.34)	0.379 (0.47, 0.37, 0.30)	0.384
World model	0.315 (0.24, 0.33, 0.37)	0.299 (0.25, 0.28, 0.37)	0.050

Shortlist product.

The operational output is a ranked list of OD corridors. Table 5 shows the top 20 of the seed-0 frozen-holdout pool, ranked by the tabular pathway. The pool contains 27 pairs: 13 held-out observed corridors (“new-corridor positives”) and 14 hard-matched unlabeled pairs. Twelve of the 13 held-out positives rank in the top 20, against about 9.6 expected for a random ordering of the pool; the remaining positive (Romania–Serbia) ranks lower. Unlabeled pairs in the list are sampled non-observed dyads and should be read as candidates for review, not as claimed routes. World-model scores are nearly constant across the list (0.458–0.473) and are not used for ranking. Scores here come from a repeat run under the same protocol and differ slightly from the reported run (average precision 0.73 versus 0.78).

8 Ethical considerations

Rankings of trafficking corridors can be misused, and models trained on detection-conditioned data can inherit reporting bias (Lum and Isaac 2016). The data used here are publicly available and aggregated at the country-pair level, and the outputs contain no victim-level information. The rankings are intended to help prioritize corridors for review, and any operational use should include human review and attention to the cost of false positives.

Table 5 Top 20 corridors of the seed-0 frozen-holdout pool, ranked by tabular score. “Held-out positive” is an observed CTDC corridor withheld from training; “unlabeled” is a sampled non-observed pair. Not externally validated.

Rank	Origin	Destination	Role	Score
1	Moldova	Macedonia	unlabeled	0.716
2	Cambodia	Thailand	held-out positive	0.540
3	Korea, South	United States	held-out positive	0.488
4	Indonesia	United Arab Emirates	held-out positive	0.485
5	Ukraine	United Arab Emirates	held-out positive	0.480
6	Kyrgyzstan	Kazakhstan	held-out positive	0.438
7	Vietnam	Malaysia	held-out positive	0.419
8	Colombia	Ecuador	held-out positive	0.411
9	Belarus	Bahrain	unlabeled	0.372
10	Kenya	Saudi Arabia	held-out positive	0.365
11	Belarus	Israel	unlabeled	0.353
12	Philippines	Japan	held-out positive	0.329
13	Burkina Faso	Senegal	held-out positive	0.315
14	Ukraine	Romania	unlabeled	0.293
15	Ukraine	Cyprus	unlabeled	0.280
16	Thailand	Indonesia	held-out positive	0.269
17	Ukraine	Austria	held-out positive	0.256
18	Romania	Hungary	unlabeled	0.238
19	Sri Lanka	Armenia	unlabeled	0.225
20	Moldova	Finland	unlabeled	0.225

9 Discussion

Systems interpretation.

The engineered response to partial OD observability is a verified ranking loop: representation constraints, multi-pathway inference, and intervention-based assurance. The shortlist couples computational inference to human or institutional review. The value of the loop is as much in what it reveals as in what it passes: the same battery that supports the label-use and null-model checks also exposes unstable rewiring, weak world-model interventions, and an inert future-route metric.

<i>Relational channel.</i>	737
	738
GNNs implement network-science link ranking with message passing (Liben-Nowell	739
and Kleinberg 2007 ; Zhang and Chen 2018). At CTDC scale (74 positives) they do	740
not dominate tabular attributes, and rewire assurance is seed-unstable. This mirrors	741
broader cautions that graph-model gains depend on evaluation design (Shchur et al.	742
2018 ; Li et al. 2023). Graph representation remains the natural state space for OD	743
corridors; trust in the relational pathway should track assurance outcomes.	744
	745
	746
	747
	748
	749
<i>Limits.</i>	750
	751
CTDC coverage after harmonization is thin; mapping errors bound labels; feature	752
years and CTDC registration years are imperfectly aligned; HTI context is a single	753
cross-section in the reported run; test splits contain about a dozen positives, so seed	754
variance is large and pathway differences should be read as tendencies; prospective	755
cuts are evaluated on a random test slice rather than a cut-specific OD frame, and	756
only part of each prospective set falls inside the candidate frame; rewire stability needs	757
more seeds or denser OD layers; and prospective cuts are not operational ground truth.	758
The world-model future-route task collapses to the current label in this protocol and	759
is therefore uninformative.	760
	761
	762
	763
	764
	765
	766
	767
	768
	769
	770
10 Conclusion	771
	772
We presented an auditable corridor shortlisting system for partially observed traf-	773
ficking OD data. CTDC-sparse labels, hard-matched negatives, and coupled tabular,	774
GNN, and world-model rankers produce a top- K corridor list; an assurance battery	775
and prospective CTDC cuts verify properties of that list before external review. Multi-	776
seed results support label-use and null-model checks, while showing that graph-rewire	777
dependence is unstable, world-model interventions are weak, and the tabular advan-	778
tage in prospective recovery is driven by one seed. The contribution is a reproducible	779
	780
	781
	782

783 ranking and verification loop, not a claim that shortlisted corridors are confirmed
784 trafficking routes.
785

786

787

788 **List of abbreviations**

789

790 *AUC-ROC* area under the receiver operating characteristic curve

791

792 *CTDC* Counter-Trafficking Data Collaborative

793

794 *ECE* expected calibration error

795

796 *GNN* graph neural network

797

798 *GRU* gated recurrent unit

799

800 *HTI* Human Trafficking Indicators

801

802 *ISCS* International Symposium on Complex Systems

803

804 *kNN* *k*-nearest neighbor

805

806 *MLP* multilayer perceptron

807

808 *OD* origin–destination

809

810 *PR-AUC* area under the precision–recall curve

811

812

813 **Declarations**

814

815 *Use of large language models.*

816

817 A large language model (Claude, Anthropic) was used to assist with reformatting

818

819 the manuscript for submission. Cursor was used to assist with coding. The authors

820

821 reviewed and verified all content and take full responsibility for it.

822

823 *Ethics approval and consent to participate.*

824

825 Not applicable.

826

827 *Consent for publication.*

828

829 Not applicable.

830

<i>Availability of data and materials.</i>	829
The datasets analysed during the current study are publicly available: the Counter-	830
Trafficking Data Collaborative global dataset (Counter-Trafficking Data Collaborative 2017)	831
and the Human Trafficking Indicators dataset (Frank 2022). The	832
datasets generated during the current study (per-seed falsification and metrics tables,	833
prospective-recovery tables, and scored shortlists) are available upon request.	834
	835
	836
	837
	838
	839
<i>Competing interests.</i>	840
The authors declare that they have no competing interests.	841
	842
	843
	844
<i>Funding.</i>	845
This research received no funding.	846
	847
	848
	849
<i>Authors' contributions.</i>	850
All authors contributed to the conception of the study. MAM developed the code, ran	851
the experiments, analysed the results, and wrote and edited the manuscript. MZE,	852
DC, and TL reviewed the manuscript and approved the final version.	853
	854
	855
	856
<i>Acknowledgements.</i>	857
Not applicable.	858
	859
	860
	861
References	862
	863
Arroyave FJ, Petersen AM, Jenkins J, et al (2020) Multiplex networks reveal geo-	864
graphic constraints on illicit wildlife trafficking. Appl Netw Sci 5:20. https://doi.	865
org/10.1007/s41109-020-00262-6	866
	867
	868
	869
Beine M, Bertoli S, Fernández-Huertas Moraga J (2016) A practitioners' guide to	870
gravity models of international migration. World Econ 39(4):496–512. https://doi.	871
org/10.1111/twec.12265	872
	873
	874

875 Bekker J, Davis J (2020) Learning from positive and unlabeled data: a survey. *Mach*
 876 *Learn* 109(4):719–760. <https://doi.org/10.1007/s10994-020-05877-5>
 877
 878
 879 Beyer RM, Schewe J, Lotze-Campen H (2022) Gravity models do not explain, and
 880 cannot predict, international migration dynamics. *Humanit Soc Sci Commun* 9:56.
 881 <https://doi.org/10.1057/s41599-022-01067-x>
 882
 883
 884
 885 Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction
 886 of missing links in networks. *Nature* 453(7191):98–101. [https://doi.org/10.1038/](https://doi.org/10.1038/nature06830)
 887 [nature06830](https://doi.org/10.1038/nature06830)
 888
 889
 890
 891 Counter-Trafficking Data Collaborative (2017) Global data hub on human trafficking.
 892 Launched 2017 by the International Organization for Migration, Polaris and Liberty
 893 Asia, URL <https://www.ctdatacollaborative.org>, accessed [TODO: access date]
 894
 895
 896
 897 Dimas GL, Konrad RA, Maass KL, et al (2022) Operations research and analytics to
 898 combat human trafficking: A systematic review of academic literature. *PLoS One*
 899 17(8):e0273708. <https://doi.org/10.1371/journal.pone.0273708>
 900
 901
 902
 903 Feingold DA (2010) Trafficking in numbers: The social construction of human traf-
 904 ficking data. In: Andreas P, Greenhill KM (eds) *Sex, Drugs, and Body Counts: The*
 905 *Politics of Numbers in Global Crime and Conflict*. Cornell University Press, Ithaca,
 906 NY, p 46–74, <https://doi.org/10.7591/9780801458309-005>
 907
 908
 909
 910 Fosdick BK, Larremore DB, Nishimura J, et al (2018) Configuring random graph
 911 models with fixed degree sequences. *SIAM Rev* 60(2):315–355. [https://doi.org/10.](https://doi.org/10.1137/16M1087175)
 912 [1137/16M1087175](https://doi.org/10.1137/16M1087175)
 913
 914
 915
 916 Frank RW (2022) Human trafficking indicators: A new dataset. *Int Interact* 48(1):152–
 917 171. <https://doi.org/10.1080/03050629.2021.1968387>
 918
 919
 920

Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In:	921
Proceedings of the 34th International Conference on Machine Learning, pp 1321–	922
1330	923
	924
	925
	926
Kazemi SM, Goel R, Jain K, et al (2020) Representation learning for dynamic graphs:	927
A survey. J Mach Learn Res 21(70):1–73	928
	929
	930
Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional	931
networks. In: International Conference on Learning Representations	932
	933
	934
Kivelä M, Arenas A, Barthelemy M, et al (2014) Multilayer networks. J Complex	935
Netw 2(3):203–271. https://doi.org/10.1093/comnet/cnu016	936
	937
	938
Li J, Shomer H, Mao H, et al (2023) Evaluating graph neural networks for link predic-	939
tion: Current pitfalls and new benchmarking. In: Advances in Neural Information	940
Processing Systems 36 (Datasets and Benchmarks Track)	941
	942
	943
	944
Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks.	945
J Am Soc Inf Sci Technol 58(7):1019–1031. https://doi.org/10.1002/asi.20591	946
	947
	948
Lü L, Zhou T (2011) Link prediction in complex networks: A survey. Physica A	949
390(6):1150–1170. https://doi.org/10.1016/j.physa.2010.11.027	950
	951
	952
Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19. https://doi.	953
org/10.1111/j.1740-9713.2016.00960.x	954
	955
	956
	957
Magid MA, Ertem MZ, Cingranelli D, et al (2026) Temporal graph networks for hid-	958
den route discovery in partially observed human trafficking systems. In: Cherifi H,	959
Guillaume JL, Melancon G (eds) Proceedings of the Sixth International Symposium	960
on Complex Systems (ISCS 2026), Studies in Systems, Decision and Control, vol	961
709. Springer, in press	962
	963
	964
	965
	966

967 Meneghini C, Aziani A, Dugato M (2020) Modeling the structure and dynamics of
 968 transnational illicit networks: an application to cigarette trafficking. *Appl Netw Sci*
 969 5:21. <https://doi.org/10.1007/s41109-020-00265-3>
 970
 971
 972
 973 Raji ID, Smart A, White RN, et al (2020) Closing the AI accountability gap: defining
 974 an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the*
 975 *2020 Conference on Fairness, Accountability, and Transparency*, pp 33–44, <https://doi.org/10.1145/3351095.3372873>
 976
 977
 978
 979
 980 Ramchandani P, Bastani H, Wyatt E (2025) Unmasking human trafficking risk in
 981 commercial sex supply chains with machine learning. *Manuf Serv Oper Manag*
 982 27(3):700–719. <https://doi.org/10.1287/msom.2022.0304>
 983
 984
 985
 986 Rossi E, Chamberlain B, Frasca F, et al (2020) Temporal graph networks for deep
 987 learning on dynamic graphs. *ArXiv preprint arXiv:2006.10637*, [arXiv:2006.10637](https://arxiv.org/abs/2006.10637)
 988
 989
 990 Shchur O, Mumme M, Bojchevski A, et al (2018) Pitfalls of graph neural network
 991 evaluation. *ArXiv preprint arXiv:1811.05868*, [arXiv:1811.05868](https://arxiv.org/abs/1811.05868)
 992
 993
 994 Shelley L (2014) Human smuggling and trafficking into Europe: A
 995 comparative perspective. Tech. rep., Migration Policy Institute,
 996 Washington, DC, URL [https://www.migrationpolicy.org/research/](https://www.migrationpolicy.org/research/human-smuggling-and-trafficking-europe-comparative-perspective)
 997
 998
 999 [human-smuggling-and-trafficking-europe-comparative-perspective](https://www.migrationpolicy.org/research/human-smuggling-and-trafficking-europe-comparative-perspective), accessed 1 Oct
 1000
 1001 2026
 1002
 1003 Silverman BW (2020) Multiple-systems analysis for the quantification of modern slav-
 1004 ery: classical and Bayesian approaches. *J R Stat Soc Ser A Stat Soc* 183(3):691–736.
 1005
 1006 <https://doi.org/10.1111/rssa.12505>
 1007
 1008
 1009 Tyldum G (2010) Limitations in research on human trafficking. *Int Migr* 48(5):1–13.
 1010
 1011 <https://doi.org/10.1111/j.1468-2435.2009.00597.x>
 1012

Veličković P, Cucurull G, Casanova A, et al (2018) Graph attention networks. In:	1013
International Conference on Learning Representations	1014
	1015
	1016
Weitzer R (2014) New directions in research on human trafficking. Ann Am Acad Pol	1017
Soc Sci 653(1):6–24. https://doi.org/10.1177/0002716214521562	1018
	1019
	1020
Zhang M, Chen Y (2018) Link prediction based on graph neural networks. In: Advances	1021
in Neural Information Processing Systems	1022
	1023
	1024
	1025
	1026
	1027
	1028
	1029
	1030
	1031
	1032
	1033
	1034
	1035
	1036
	1037
	1038
	1039
	1040
	1041
	1042
	1043
	1044
	1045
	1046
	1047
	1048
	1049
	1050
	1051
	1052
	1053
	1054
	1055
	1056
	1057
	1058