SynTraffic: A synthetic dataset for spatiotemporal human trafficking detection in heterogeneous graphs

João Lopes[1,2], Ana Nascimento[1], Pratheeksha Nair[3,4], Catalina Vajiac[5], Andreas Olligschlaeger[6], Reihaneh Rabbany[3,4], Christos Faloutsos[5], Mirela Cazzolato[1]

[1] ICMC - University of São Paulo (USP), [2] FFCLRP - University of São Paulo (USP)
[3] McGuill University, [4] Mila - Quebec AI Institute, [5] Carnegie Mellon University, [6] i3 LLC

joaovlpss@usp.br, mirela@icmc.usp.br

Abstract. *Online escort platforms have become a common vector for human trafficking (HT) networks to advertise victims alongside legitimate listings, making manual screening infeasible. Progress on automated analysis is limited by restricted data access: HT advertisement corpora carry personal, potentially victim-identifying information, imposing ethical and legal constraints on sharing and reproducibility. This work addresses that gap with a reproducible methodology for generating synthetic HT advertisement data that is statistically representative of a real corpus, without exposing sensitive content. The proposed dataset SynTraffic reproduces posting volume, spatial mobility, temporal dynamics, and metadata-sharing structure, including the size distribution of identity clusters formed by shared contact metadata. We statistically validate SynTraffic testing distributional fidelity and cluster-size fidelity. SynTraffic is openly available, as well as the the generation methodology, presenting a privacy-safe substitute for the real corpus.*

1. Introduction

Human trafficking (HT) constitutes a severe violation of human rights. The International Labour Organization estimated that 27.6 million people were living in forced labor conditions on any given day in 2021 [International Labour Organization et al. 2022], with women and girls disproportionately affected by forced sexual exploitation. Criminal networks exploit online escort platforms to advertise victims, embedding illicit content among legitimate and spam postings and leveraging the anonymity and scale of these environments to evade detection and operate across cities and borders [Nair et al. 2024]. With thousands of advertisements posted daily, manual investigation becomes impractical and motivates Data Science and AI approaches to detect suspicious patterns at scale. Heterogeneous graphs, in particular, have proven effective for detecting malicious accounts on large platforms [Liu et al. 2018].

A recurring signal in this domain is coordinated mobility, in which victim groups are advertised across multiple cities over time, distinguishing organized trafficking from independent activity. Prior work has tackled this problem via information-theoretic clustering, unsupervised text mining, visual analytics, and graph-based learning [Nagpal et al. 2017, Vajiac et al. 2023b, Kulshrestha 2021, Nair et al. 2024]. These approaches have advanced the field, yet one obstacle remains central to reproducibility: HT datasets carry personal victim information, imposing strict ethical and legal

constraints on sharing and reuse. Most published methods therefore rely on proprietary corpora that cannot be released, a limitation the community openly acknowledges [Lee et al. 2021, Vajiac et al. 2023a].

This work proposes a reproducible methodology for generating synthetic online advertisement data that is statistically representative of a real HT corpus. We propose the SynTraffic dataset, whose marginal and joint distributions of posting volume, spatial mobility, temporal patterns, and metadata-sharing structure closely match a real-world corpus. We remove the textual content, since generating realistic text would require formal privacy guarantees beyond this work's scope [Beduschi 2024].

The generation process is guided by Exploratory Data Analysis (EDA) over the real corpus, and statistically validated with tests of distributional fidelity and cluster-size fidelity per metadata field. SynTraffic is not proposed as a HT detection benchmark: the real corpus carries no ground-truth label distinguishing trafficking from independent or legitimate postings. In fact, this is one of the main challenges in this domain: *how can we detect organized, suspicious activities?* SynTraffic is instead a statistically faithful, privacy-safe substitute for the real corpus, suited to developing and testing data-processing pipelines without exposing sensitive data. SynTraffic and the generation methodology are openly available[1].

2. Preliminary concepts and related work

Synthetic data generation offers a way to share research artifacts under privacy constraints. Given a sensitive dataset, the goal is a synthetic corpus whose marginal and joint statistical distributions closely mirror the original without retaining any individual record, a common practice in healthcare, criminal justice, and human trafficking research [Beduschi 2024]. This trades off fidelity against privacy. We takle this problem by synthesizing only structural and spatiotemporal metadata and excluding free-text fields.

Prior HT detection work has approached the problem from complementary angles, summarized in Table 1. Early studies framed it as entity resolution, linking ads through shared phone numbers and metadata [Nagpal et al. 2017]. InfoShield [Lee et al. 2021] and DeltaShield [Vajiac et al. 2023b] cluster near-duplicate ads via information-theoretic compression. TRAFFICLight [Kulshrestha 2021] applies unsupervised learning over textual features. TrafficVis [Vajiac et al. 2023a] offers an interactive visual analytics interface. T-NET [Nair et al. 2024] casts detection as weakly supervised graph learning over shared metadata. Finally, in [Summers et al. 2023] the authors combine textual and image features in multimodal classifiers. These methods are validated almost exclusively on proprietary data, motivating our focus on a reproducible, shareable alternative.

3. Synthetic data generation and the SynTraffic dataset

The real dataset contains 26,452 posts from an online escort advertisement website, with 4,927 unique users, covering user ID, posting timestamp, locality, latitude, longitude, phone number, e-mail, social media handle, WhatsApp, image URL, and contact URL (see attribute description in Table 2). Textual fields and country were excluded from synthesis. The remaining attributes preserve the original schema. Generation was guided by successive EDA over the real dataset, whose findings motivated each modeling decision.

[1] *https://github.com/joaovlpss/synthetic_ht_dataset*

Table 1. Compact taxonomy of related work on human trafficking detection. Specification: ✓ = coverage; ✗ = absence.

Work	Year	Input			Dataset	
		Text	Meta.	Traj.	Public	Synth.
[Nagpal et al. 2017]	2017	✗	✓	✗	✗	✗
[Lee et al. 2021]	2021	✓	✗	✗	✗	✗
[Kulshrestha 2021]	2021	✓	✓	✗	✗	✗
[Vajiac et al. 2023a]	2023	✓	✓	✗	✗	✗
[Vajiac et al. 2023b]	2023	✓	✗	✗	✗	✗
[Summers et al. 2023]	2023	✓	✗	✗	✗	✗
[Nair et al. 2024]	2024	✓	✓	✗	✗	✓
Ours (SynTraffic)	**2026**	✗	✓	✓	✓	✓

Table 2. Attributes of the advertisement dataset.

Attribute	Description
User identifier	Display name of the post author on the platform.
Posting timestamp	Exact date and time at which the advertisement was published.
Locality	Name of the city or municipality where the service is advertised.
Latitude-longitude	Geographic coordinates of the posting locality.
Phone number	Phone number provided by the user in the advertisement.
Email address	Email address provided by the user in the advertisement.
Social media handle	External identifier on websites or social networks provided by the user.
WhatsApp	WhatsApp number provided by the user in the advertisement.
Main image URL	URL of the featured image of the advertisement.
Contact URL	External contact address associated with the user.

A. Posting volume per user. For each real user p, we compute the post count $c_p = |\{a : \text{author}(a) = p\}|$. The distribution has the long-tail behavior characteristic of this domain: most accounts publish few advertisements, while a minority concentrate volumes well above the median, a signal that distinguishes individual accounts from organized operations or spam. A Gaussian Kernel Density Estimate (KDE) is fitted over the real count, from which N values are sampled to define each synthetic user's posting volume.

B. Behavioral profiles and geolocation. Users are grouped into four profiles according to their number of posts c_p, using breakpoints $1, 10, 100, \infty$: *unique*, *casual*, *frequent*, and *massive*. For each profile p, we derive empirical distributions of the number of distinct localities ($\mathcal{L}_p$) and posting-hour dispersion ($\mathcal{S}_p$). High-volume profiles exhibit greater spatial mobility and lower temporal dispersion (Figure 1).

For each synthetic user, the number of localities and temporal dispersion are sampled from $\mathcal{L}_p$ and $\mathcal{S}_p$, respectively. An anchor hour μ is drawn from the global bimodal posting-hour distribution $\mathcal{H}$, and posting hours are sampled around μ using Gaussian noise with the sampled dispersion. The user's locality basket is sampled without replacement, weighted by real posting frequency.

Location names are sampled separately from their geospatial attributes (latitude, longitude, and region), which are subsequently assigned through a per-location lookup.

This preserves the real data's structural sparsity: some named localities are consistently ungeocoded, and joint sampling would exclude them from generation entirely.

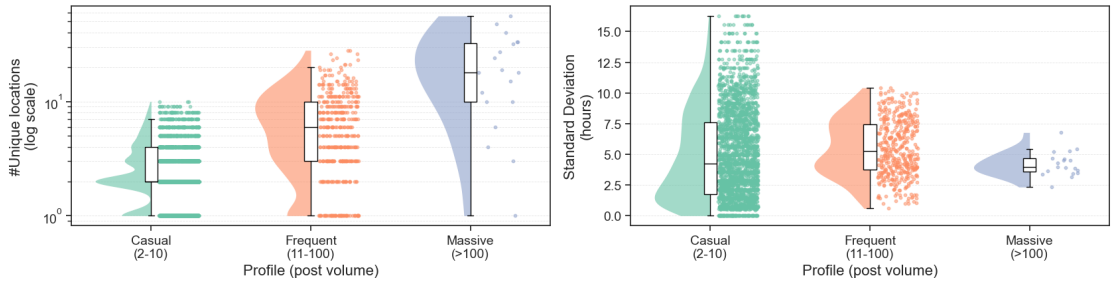

Figure 1. Raincloud plots showing the distribution of unique locations (log scale) and temporal dispersion by user profile, combining density, raw data points, and summary statistics.

C. Posting dates and user lifespan. We characterize each user's *lifespan* (days between first and last post) and mean inter-arrival time (IAT, difference in days between consecutive posts), both strongly profile-dependent: heavy users have short IATs and long lifespans. This motivates modeling the temporal sequence as a random walk: an empirical increment pool $\mathbb{D}_p$ (real day-deltas between consecutive posts, by profile) is used to (i) draw $c_u - 1$ increments and accumulate them, (ii) select a start offset within the global window $[t_{\min}, t_{\max}]$ so the sequence fits, and (iii) reconcile the resulting day with the modeled hour and a random minute offset.

D. Metadata in identity clusters. Metadata sharing across users is a crucial characteristic of HT data, which can indicate coordinated *modus operandi*. We model four dynamic fields: phone, email, social, WhatsApp. Ad and image URLs are generated uniquely per user, as sharing is rare. Three findings guide the generator. **(F1) sharing cardinality:** most values are linked to a single account, while a long tail is shared by multiple accounts. **(F2) sparsity and consistency:** fields are often absent and inconsistently used across posts, motivating a two-level model with binary *ownership* and a user-specific probabilistic *usage* rate. **(F3) co-occurrence:** ownership indicators correlate across fields (*e.g.*, WhatsApp with phone), requiring joint rather than independent field assignment.

The generation proceeds in two independent stages. In the **ownership stage**, each synthetic user is assigned a presence signature, defined as a boolean tuple of which fields they own, sampled from the empirical distribution of real per-user signatures, preserving F3. In the **sharing stage**, each field is treated independently: the subset of users who own it is partitioned into value-sharing groups, sized from that field's own empirical sharing-cardinality distribution (F1), and each group receives one globally unique *Faker*-generated value. This independence is necessary because real sharing-cardinality distributions differ sharply by field (phone groups reach 24 users, while over 98% of emails belong to a single user) and an earlier design reusing one partition (sized by phone) across all fields was found, during validation, to distort every other field's size distribution. In the **usage stage**, a profile-conditioned dropout applies: for each owned field, a rate r is sampled from the real rate pool of the user's profile and field, and each post retains the value with probability r. This design assigns one value per field per user and does not model within-user rotation over time: 39.5% of real users are associated with more than one phone number across their history (mean 1.97, max 68), which the generator does

not reproduce. This is the main limitation observed, and it accounts for an undercount in the absolute number of phone-sharing groups (Section 4), even though the group-size distribution's shape is statistically indistinguishable from the real one.

4. Statistical Validation

We validate SynTraffic with quantitative statistical tests, detailed next.

Distributional fidelity. We assess distributional fidelity using the SDMetrics (Synthetic Data Metrics) quality-report framework. *Column Shapes* measures univariate fidelity using the Kolmogorov-Smirnov complement (KS-complement) for continuous attributes and the Total Variation Distance complement (TV-complement) for categorical and Boolean attributes. *Column Pair Trends* measures bivariate fidelity based on contingency and correlation similarity. Both scores range from 0 to 1, where 1 indicates statistical indistinguishability. Applied to the behavioral attributes and each dynamic field's ownership indicator (Table 3), the overall quality score is 95.25% (*Column Shapes*: 96.42%; *Column Pair Trends*: 94.07%).

Metadata sharing-cardinality fidelity. Sharing cardinality is a cross-row relation. For each dynamic field we compare real and synthetic group-size distributions via a two-sample KS test (Table 4). All four fields are statistically indistinguishable in shape (all $p > 0.89$). Phone is the only field where the absolute group count differs substantially (2,102 vs. 4,123), attributable to the limitation noted in Section 3.D.

Table 3. SDMetrics distributional fidelity per attribute (Score 1.0 indicates statistical indistinguishability).

Attribute	Metric	Score	Attribute	Metric	Score
location_detail	TVD	0.9165	longitude	KS	0.9677
has_social	TVD	0.9315	latitude	KS	0.9700
data_region_id	TVD	0.9537	has_phone	TVD	0.9812
has_whatsapp	TVD	0.9950	has_email	TVD	0.9983

TVD (Total Variation Distance): TVComplement; KS (Kolmogorov-Smirnov): KSComplement.

Table 4. Metadata sharing-cardinality fidelity: two-sample KS test between real and synthetic group-size distributions.

Field	Real groups	Synth. groups	Real max	Synth. max	KS p-value
phone	4,123	2,102	24	16	0.9999
email	1,464	1,430	4	4	1.0000
social	378	188	16	15	0.9999
whatsapp	938	688	16	15	0.8935

5. Conclusion

This paper presented a reproducible methodology for generating synthetic HT ads data that is statistically representative of a real-world corpus, addressing the restricted data access that limits reproducibility in this field. SynTraffic is calibrated on empirical distributions from a real corpus and reproduces its posting volume, spatial mobility, temporal patterns, and metadata-sharing structure, validated through quantitative statistical tests.

SynTraffic is proposed as a statistically faithful, privacy-safe substitute for the real corpus, suited to developing and testing data-processing pipelines. It is not proposed as a detection benchmark, since the real corpus carries no ground-truth label distinguishing HT from independent or legitimate postings. Future work will explore heterogeneous graph representations and spatiotemporal trajectory features built on SynTraffic, along with extensions to the generation methodology.

Acknowledgement. This work was supported by the São Paulo Research Foundation (FAPESP – grants #2026/11130-2, #2024/15430-5, #2024/13328-9), CAPES – Finance Code 001, CNPq – grants #401496/2025-2 and #402987/2025-0, and the University of São Paulo (PRPI 1032 - grant #207, and PUB - grants #4622 and #5192).

References

Beduschi, A. (2024). Synthetic data protection: Towards a paradigm change in data regulation? *Big Data &; Society*, 11(1). DOI: 10.1177/20539517241231277.

International Labour Organization, Walk Free, and International Organization for Migration (2022). Global estimates of modern slavery: Forced labour and forced marriage. Technical report, International Labour Organization, Geneva. Accessed: 21 May 2026.

Kulshrestha, A. (2021). Detection of organized activity in online escort advertisements. Master's thesis, McGill University, Montreal, Canada. URL: https://escholarship.mcgill.ca/concern/theses/6395wd28p.

Lee, M.-C., Vajiac, C., Kulshrestha, A., Levy, S., Park, N., Jones, C., Rabbany, R., and Faloutsos, C. (2021). InfoShield: Generalizable information-theoretic human-trafficking detection. In *IEEE ICDE*. DOI: 10.1109/ICDE51399.2021.00101.

Liu, Z., Chen, C., Yang, X., Zhou, J., Li, X., and Song, L. (2018). Heterogeneous graph neural networks for malicious account detection. In *ACM CIKM*, page 2077–2085, New York, NY, USA. DOI: 10.1145/3269206.3272010.

Nagpal, C., Miller, K., Boecking, B., and Dubrawski, A. (2017). An entity resolution approach to isolate instances of human trafficking online. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 77–84, Copenhagen, Denmark. Association for Computational Linguistics. DOI: 10.18653/v1/w17-4411.

Nair, P., Liu, J., Vajiac, C., Olligschlaeger, A., Chau, D. H., Cazzolato, M., Jones, C., Faloutsos, C., and Rabbany, R. (2024). T-net: weakly supervised graph learning for combatting human trafficking. AAAI. DOI: 10.1609/aaai.v38i20.30233.

Summers, L., Shallenberger, A. N., Cruz, J., and Fulton, L. V. (2023). A multi-input machine learning approach to classifying sex trafficking from online escort advertisements. *Machine Learning and Knowledge Extraction*, 5(2):460–472.

Vajiac, C., Chau, D. H., Olligschlaeger, A. M., Mackenzie, R., Nair, P., Lee, M., Li, Y., Park, N., Rabbany, R., and Faloutsos, C. (2023a). Trafficvis: Visualizing organized activity and spatio-temporal patterns for detecting and labeling human trafficking. *IEEE Trans. Vis. Comput. Graph.*, 29(1):53–62. DOI: 10.1109/TVCG.2022.3209403.

Vajiac, C., Lee, M.-C., Kulshrestha, A., Levy, S., Park, N., Olligschlaeger, A., Jones, C., Rabbany, R., and Faloutsos, C. (2023b). Deltashield: Information theory for human-trafficking detection. *ACM Trans. Knowl. Discov. Data*, 17(2). DOI: 10.1145/3563040.