Unmasking human trafficking risk in commercial sex supply chains with machine learning

P. Ramchandani[*,1], H. Bastani[*,1], E. Wyatt[2]

[1] Department of Operations, Information and Decisions, Wharton School, University of Pennsylvania

[2] Uncharted Software, Founding Partner of the TellFinder Alliance

[*]Corresponding authors. Email: piar2@wharton.upenn.edu, hamsab@wharton.upenn.edu

The covert nature of sex trafficking provides a significant barrier to generating large-scale, data-driven insights to inform law enforcement, policy and social work. We leverage massive deep web data (collected globally from leading commercial sex websites) in tandem with a novel machine learning framework to unmask suspicious recruitment-to-sales pathways, thereby providing the first global network view of trafficking risk in commercial sex supply chains. This allows us to infer likely recruitment-to-sales trafficking routes of criminal entities, deceptive approaches used to recruit victims, and regional variations in recruitment vs. sales pressure. These insights can help law enforcement agencies along trafficking routes better coordinate efforts, as well as target local counter-trafficking policies and interventions towards exploitative behavior frequently exhibited in that region.

Summary: We leverage large-scale deep web data and machine learning to provide the first global view of trafficking risk in commercial sex supply chains, from recruitment to sales.

According to the FBI, sex trafficking is the fastest growing organized crime business and the third largest criminal enterprise in the world [1]. The International Labor Organization estimates there were 4.8 million sex trafficking victims in 2017 alone [2]. Consequently, there is high demand from field experts [3, 4, 5] and academics [6, 7, 8, 9, 10] for a *large-scale* and *data-driven* view of the underlying supply chain dynamics [11] of trafficking that can inform law enforcement, policy and social work. For instance, understanding where and how victims are recruited in different regions can enable *preventative* interventions at the source of the supply chain (recruitment) in contrast to prevalent mitigation strategies that target the end of the supply chain (sales) [12, 13]. Furthermore, inferring likely recruitment-to-sales trafficking routes of criminal entities can enhance coordination strategies between relevant law enforcement agencies and task forces to increase efficiency of counter-trafficking efforts [14, 15, 16, 17, 18, 19, 12].

However, the covert nature of trafficking provides a significant barrier to generating such insights. For example, limited existing research literature on sex trafficking uses any data, and those that do primarily leverage qualitative interviews with trafficking survivors [20]. It is hard to generate quantitative and generalizable insights from such interviews, because they are

qualitative in nature and severely limited in scale; moreover, they can be traumatic for victims and can result in unreliable information [7].

In this paper, we propose to leverage unstructured, massive *deep web data* to characterize sex trafficking recruitment and sales risk at scale. The deep web represents portions of the World Wide Web that are not indexed by traditional search engines, e.g., temporary or dynamic content from private websites that can only be accessed via specialized queries. A significant portion of commercial sex activity – and the exploitative behavior that accompanies it – occurs online [21, 22], making the deep web a rich and relevant data source. Trafficking is commonly targeted at vulnerable populations (e.g., 1 out of 5 homeless youth in top cities in the United States and Canada have been identified as victims of human trafficking [13]), who are frequently recruited online through "fishing" strategies that offer well-paid jobs to attract potential victims to make initial contact with traffickers [23].

We begin by leveraging data from leading commercial sex advertisement websites in conjunction with a novel machine learning framework to construct the first global view of commercial sex *supply chains*, from recruitment to sales. Importantly, however, commercial sex and sex trafficking are not synonymous [24]: "Unlawful commercial sex acts overlap with sex trafficking when participation occurs by means of force, fraud, or coercion…" [25]. In other words, it is critical to understand *how* victims are recruited into the commercial sex supply chain to distinguish trafficking victims and commercial sex workers. To address this challenge, we study how and where recruitment occurs in the supply chains we uncover; in particular, linking deceptive (non-sex) recruitment offers to commercial sex sales by the same entities (see, e.g., right panel of Figure 3A) allows us to characterize trafficking risk. Specifically, if an entity recruits victims through non-sex offers (e.g., purportedly for modeling or massage) and is also involved in commercial sex sales, then this is an indicator that trafficking may have occurred.

Identifying recruitment content in ads has historically been a significant hurdle due to the nature of sex trafficking: while sex sales ads are prevalent and convey clear intent to consumers, recruitment ads are sparse and are typically designed to trick potential victims into being trafficked. Thus, while recent work has developed techniques to scrape deep web data, extract relevant meta data (e.g., phone numbers, email addresses) and convert it into databases that support trafficking investigation inquiries by law enforcement agencies [26, 27, 28], such data has not been used for large-scale analysis of commercial sex supply chains, primarily due to the difficulty in identifying recruitment from unstructured text. We address this challenge through a novel machine learning framework that combines natural language processing, active learning, and domain expertise to distinguish recruitment and sales content at scale. We then leverage shared meta data to infer trafficking risk in commercial sex supply chain networks.

Our results yield substantial insights into the structure of commercial sex supply chains, including several policy-relevant insights. First, while sex sales predominantly occur in large

urban centers, we find evidence that recruitment is concentrated in suburban, economically constrained areas. Furthermore, there is significant variation in *how* vulnerable populations are recruited in different locations, suggesting opportunities for targeted job search training [13]. By highlighting links between deceptive (non-sex) recruitment offers and sex sales made by the same entity, we are uniquely able to infer likely trafficking routes between cities. Importantly, these routes can inform coordination strategies between relevant law enforcement agencies.

Machine Learning Framework

Our analysis is based on data supplied by the TellFinder Alliance, a network of law enforcement, technology, research, and nonprofit partners focused on adapting analytics and deep web data for counter-human trafficking applications [29]. This dataset includes approximately 14 million public, English-language commercial sex advertisements (which we call *posts*) collected from adult services websites in the deep web over a 9-month period with global geographic coverage. Nearly all posts are related to commercial sex sales, while a small subset are related to recruitment for different types of jobs (e.g., modeling, massage, etc.). Moreover, each post is associated with meta data (i.e., any extracted phone numbers, email addresses, social media handles, etc.), which can be used to connect pairs of posts that are made by the same entity. Dataset details are provided in Section 2 of the Appendix.

Our first step is to infer the underlying commercial sex supply chain. To this end, we train a deep neural network that distinguishes recruitment from sales posts based on the unstructured text in that post. A priori, all posts are unlabeled. Labels must be obtained by having a domain expert manually read the content of each post and assign a label (recruitment vs. sales); recruitment posts are additionally categorized into types (e.g., sugar parent) based on the type of employment offer made. Manually labeling all 14 million posts is clearly infeasible;[1] instead, we design an active learning approach to train a model with as few labels as possible. We face two challenges:

1. **Extreme Data Imbalance:** We estimate that only 0.06% of posts are recruitment-related, while the rest are sales, i.e., one would have to manually label nearly 2000 randomly chosen posts to find a *single* instance of recruitment in expectation. Thus, traditional supervised or active learning techniques, which rely on an initial well-balanced training set, are infeasible (see Section 1 of the Appendix for a detailed discussion).
2. **Objective Mismatch:** We seek to identify different recruitment approaches across many locations. For instance, one auxiliary task is to identify pairs of posts (one recruitment and one sales) in different locations that are linked to the same entity by their meta data; such a pair corresponds to a potential edge in the supply chain network. Thus, traditional active learning techniques that focus purely on overall accuracy may be insufficient.

[1] The private, sensitive nature of the data precludes crowdsourcing.

We leverage weak learners [30, 31] in conjunction with active learning [32, 33] to address these challenges (see first two panels of Figure 1). We give an overview of our approach in what follows; details are relegated to Section 1 of the Appendix.

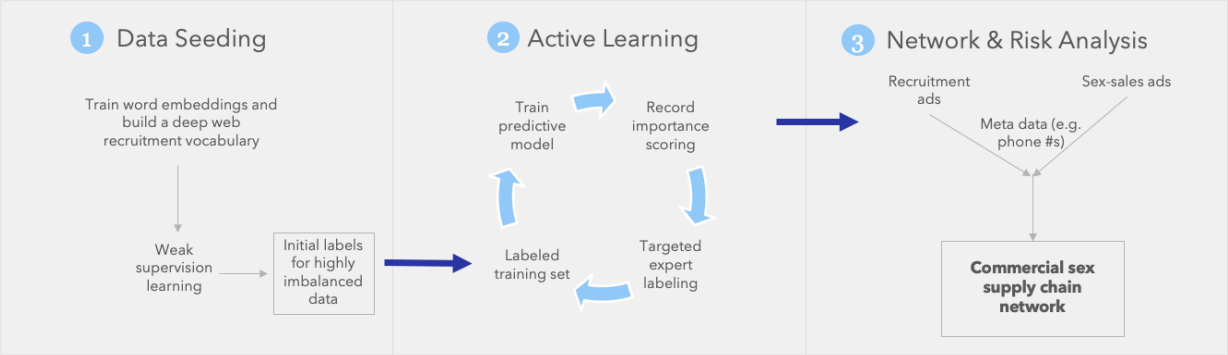

Figure 1 *Summary of our machine learning framework. We first train domain-specific word embeddings and collect expert-identified terms to develop a 'recruitment vocabulary.' This informs a weak learning heuristic to identify an initial well-balanced training set. We then apply active learning techniques (additionally incorporating geographical diversity and the likelihood of identifying new network connections) to iteratively label additional posts and update our predictive model until its performance converges. Finally, we connect recruitment and sales activity via meta data to identify supply chain networks.*

First, following standard practice, we encode unstructured text in a post into a vector space via word embeddings [34, 35]. It is well known that pre-trained embeddings perform poorly in specialized contexts [36], so we train our own domain-specific word embeddings using Gensim [37]. Next, by studying past trafficking investigations and conversing with counter-trafficking domain experts, we identify some candidate terminology that signifies recruitment risk – these include terms like "audition" or "high pay". We then develop an initial *recruitment vocabulary*, which includes all terms whose embeddings are within a short distance of the embeddings corresponding to expert-identified terms. As with traditional weak supervised learning [38, 39], the presence of a term from our recruitment vocabulary provides a noisy signal that a post may be related to recruiting. Out of our dataset of approximately 14 million posts, we identified 1651 posts that contained part of this vocabulary; we then manually labeled these posts and found that 369 of these instances corresponded to recruitment. Note that this corresponds to 22% of the labels being positive, compared to only 0.06% of the labels being positive on a random subsample of our dataset.

This process provides us with an *initial* well-balanced training set. However, it is clearly biased by the purview of domain experts and does not provide a complete view of the numerous styles/types of recruiting posts on the deep web. Thus, we use pool-based active learning, which is known to improve classifiers with significantly reduced manual labeling effort [40, 41]. Rather than labeling a random subset of posts, these approaches direct costly labeling effort towards posts that are estimated to resolve the most uncertainty (i.e., improve the accuracy) of the current classifier. In particular, we begin by training an initial deep neural network (which has shown great success in text classification tasks [42]) using the initial training set of 1651 posts. We then

use this classifier to assign a prediction probability to each unlabeled post on how likely it is to be recruiting-related – this metric captures the prediction uncertainty that is traditionally used by active learning to prioritize labeling [43].

However, as noted earlier, our active learning objective is not simply to maximize the accuracy of our classifier across all posts (which would have the consequence of focusing labeling efforts on locations with many posts), but to uncover an accurate representation of the underlying network *across* locations. We address this objective mismatch by incorporating geographical diversity and the likelihood of identifying new network connections in our learning procedure. Specifically, we add two additional metrics to our active learning objective: (i) a 'node information' score that prioritizes posts in under-sampled locations that may have additional recruitment activity, and (ii) an 'edge information' score that prioritizes posts corresponding to an under-sampled *pair* of locations (as determined by the meta data) that may represent a new inferred trafficking route. Our algorithm uses this objective to prioritize a batch of unlabeled posts for labeling. The resulting batch of labeled posts are then added to the labeled training data, and the deep learning network is re-trained. This active learning process is repeated until the model performance converges. Overall, we obtained labels on approximately 50,000 posts, identifying approximately 7000 recruiting-related posts. Despite the heavy data imbalance, this corresponds to 14% of the labels being positive. Furthermore, our active learning process allowed us to uncover 27 different types of recruiting tactics, far outperforming the initial expert-identified vocabulary which only identified 3 types of recruitment tactics.

Finally, we connect the identified recruitment and sales posts using shared meta data to determine which posts were made by the same entity (see last panel of Figure 1). Along with the locations of the posts, this allows us to identify the geographic network connections underlying commercial sex supply chains.

Results

Recruitment Activity. Figure 2 shows the global map of recruitment hotspots and the types of recruitment tactics identified in our data. Note that recruitment posts in the 'escort' category indicate potential sex work, while posts in all other categories do not indicate sex work. As shown in Figure 2A, we found significant activity primarily in the United States, Canada, Europe, India and Australia; this is likely due to our restriction to posts in the English language.[2]

We also observe significant geographic variation in the approaches used to recruit victims (see Figures 2A and 2B); the full list of recruitment types is in Table S3 in the Appendix. For example, within the United States, individuals are primarily targeted through modeling and porn offers in the Midwest, escort and adult entertainment services (e.g., strip clubs) in the East and West coasts, and even personal ads in several major cities. More globally, victims are targeted

[2] Future work can adapt our approach to other languages to improve global coverage.

primarily through porn and adult entertainment offers in Europe, and escort services in India. Early interventions for preventing exploitation of vulnerable populations have recommended 'job search' training to educate potential victims on the risks associated with responding to different types of recruitment posts [13]. These results can be used to tailor such educational programs towards the currently popular recruitment approaches in those specific locales.

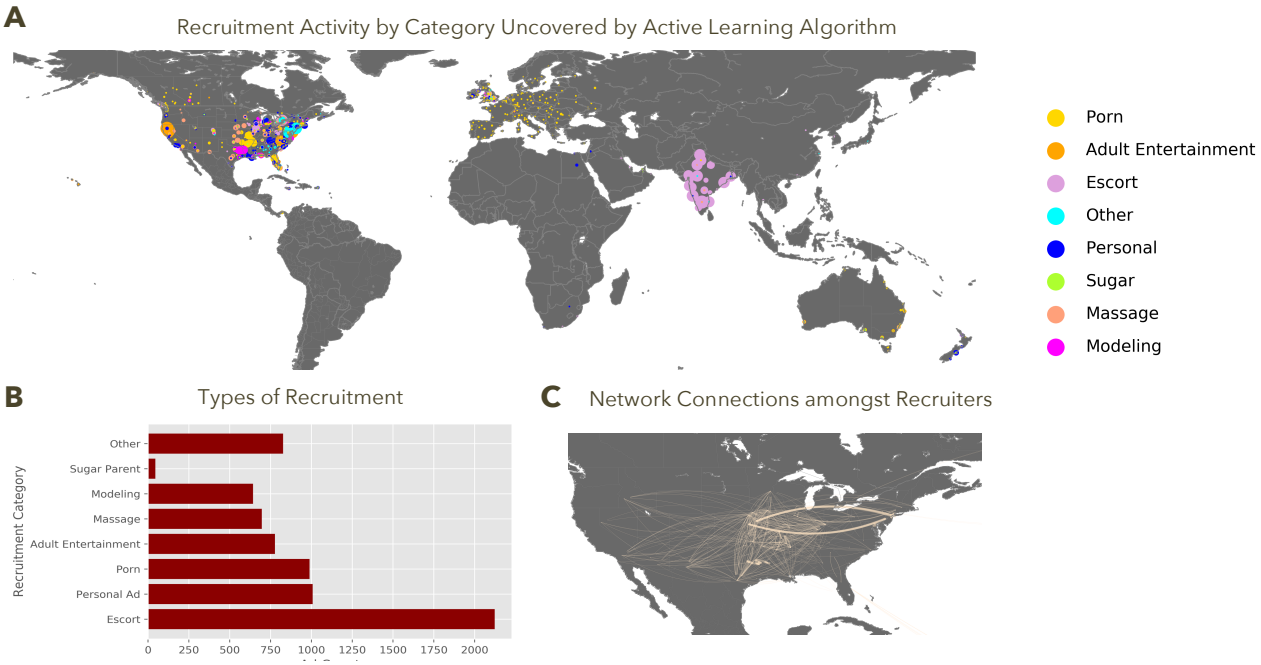

Figure 2 *The top panel (A) shows recruitment hotspots and tactics identified in our data. Larger markers indicate more posts. The bottom left panel (B) shows the histogram of recruitment posts by category across the world. The bottom right panel (C) shows the recruitment-recruitment network in the United States. We display an edge between a pair of locations if there are at least 150 recruitment posts that share meta-data (thus, are posted by the same entity); thicker lines indicate more recruitment posts (capped at 2000 posts for visual clarity).*

We also construct recruitment-recruitment networks: we create an undirected edge between a pair of locations if they each have a high volume of recruiting posts (at least 150) by the same entity. Figure 2C shows the resulting network within the United States. We observe that many recruiters operate in multiple locations spanning large distances, suggesting a highly organized effort.

Inferring Human Trafficking Risk. As discussed earlier, linking deceptive (non-sex) recruitment offers to commercial sex sales by the same entity strongly suggests that trafficking may have occurred. Figure 3A shows an example connection between an identified recruitment post and two sales posts with shared meta data; although the recruiting post offers a modeling employment opportunity in Canada, the same phone number appears in over a hundred sex sales posts in Canada, the United Kingdom, the United States and Australia. This suggests that the modeling post is a masked attempt to recruit victims into an international sex trafficking organization.

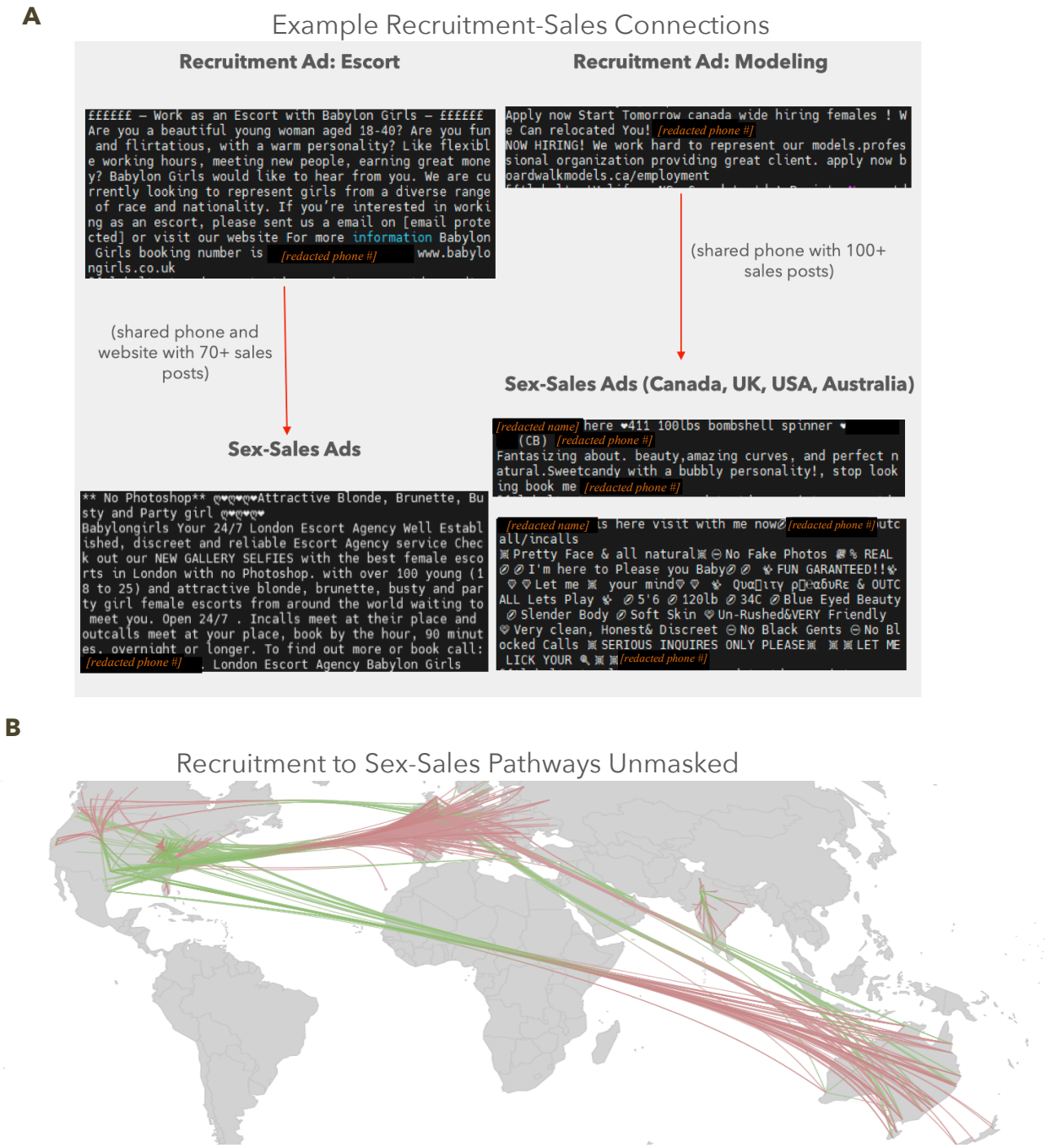

Figure 3 *The top panel (A) shows example escort and modeling recruitment posts in the UK and Canada that share the same phone number as sex sales posts in Canada and other countries; note that we have redacted personal identifiable information with square brackets and the type of information (e.g., [redacted phone #]). The bottom panel (B) shows the resulting global view of trafficking risk in commercial sex supply chain networks from deceptive recruitment offers (red) to commercial sex sales (green). Network is restricted to edges with at least 100 occurrences.*

To study trafficking risk at scale, we construct recruitment-to-sales pathways: we create a directed edge between a pair of locations if there is a recruiting post and a sales post with shared meta data. Figure 3B shows the resulting commercial sex supply chain, restricting to edges that

have at least 150 occurrences. Importantly, we find that over 95% of these connections are accounted for by deceptive recruitment posts that do not mention any potential for sex work.

We also find that 10% of recruitment ads are responsible for 85% of edges in the supply chain network, suggesting that there are a few large-scale entities driving a significant portion of trafficking activity. This result underscores the importance of our modified active learning procedure, which targets network discovery in addition to the traditional objective of improving classification accuracy.

Figure 4 delves further into the domestic recruitment-to-sales supply chain connections identified by our analysis. Domestic network connections are most prominent in the United States (Figure 4A) and India (Figure 4B).

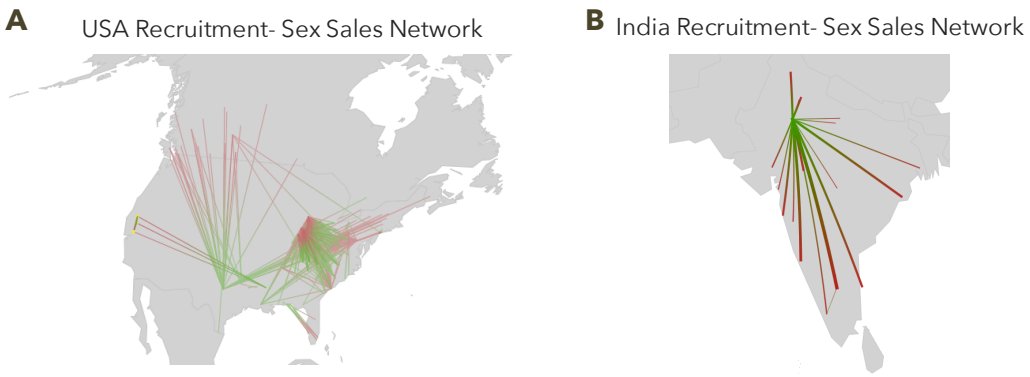

Figure 4 *Closer examination of inferred likely trafficking routes in the United States (A) and India (B). Network shows pathways from recruitment offers (red) to commercial sex sales (green) with at least 100 occurrences.*

Recruitment vs. Sales Pressure. We distinguish 'sender' cities (where victims are recruited) from 'receiver' cities (where sex sales occur). For example, in India, we observe that recruitment occurs in coastal locations, while sales primarily occur in the capital, New Delhi (see Fig. 4B). Similarly, in the United States, recruitment is concentrated in suburban locations (e.g., Scranton, Redding), while sales primarily occur in major cities (e.g., Miami, New York City, Los Angeles). Figure 5 shows a map of relative recruitment to sales pressure across the United States; we observe that densely populated locations tend to be receiver cities while less populated locations tend to be sender cities. Note that relying on traditional active learning would have directed our labeling efforts to posts from large cities (where the majority of posts occur), missing out on key recruitment hotspots in smaller cities identified by our modified active learning procedure. These results can be used to tailor interventions in specific locales, e.g., invest in education and social work to reduce recruitment in sender cities, and invest in law enforcement to prosecute sex sales in receiver cities.

We also examine the characteristics of top 50 sender and receiver cities in the United States; only 17 of these locations overlap, underscoring that recruitment and sex sales are typically

concentrated in distinct locations. Using census data, we find that sender cities tend to be more economically constrained (have higher poverty rates and lower household incomes), and furthermore have higher crime rates relative to receiver cities; details of this analysis are provided in Section 3 of the Appendix. These results suggest that sender cities may not have as many resources as larger receiver cities to prevent trafficking of their vulnerable populations; thus, they may benefit from collaborations with (better-funded) counter-trafficking agencies in larger receiver cities. Such collaborations may be particularly valuable when there is a likely recruitment-to-sales trafficking route between the two cities. For example, we identified an entity that frequently recruits (deceptively) in Redding, CA and sells sex in Sacramento, CA; therefore, a collaboration between agencies in Redding and Sacramento would simultaneously provide support for the smaller and more economically constrained Redding population, and enable targeting of a potential trafficking entity from both ends of its supply chain.

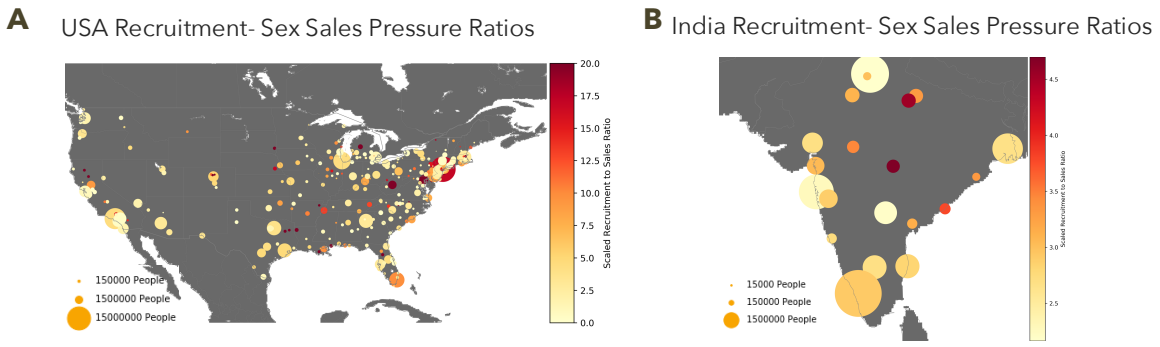

Figure 5 Map of relative recruitment to sales pressure across locations in the United States (A) and India (B). Color represents the ratio of recruitment over sales ads from the deep web scaled by a factor of 10,000 due to the substantial difference in activity levels. The size of the bubbles corresponds to population size of the city, highlighting that smaller cities tend to have higher recruitment pressure (in dark red) and larger cities have higher sex sales activity (light yellow).

Other Datasets. To the best of our knowledge, our study is the first to characterize recruitment at scale in commercial sex, which allows us to uniquely infer trafficking risk. However, we can compare our results on sex sales from the deep web against other sources. Specifically, we consider Rubmaps.ch (a popular review site for massage parlors with sexual services) as well as suspicious businesses identified through Google Places.[3] Details are provided in Section 3 of the Appendix. We find that commercial sex *sales* activity identified on the deep web roughly aligns with activity identified through other sources, but *recruitment* activity is distinct and uniquely identified by our analysis. Thus, we provide the first large-scale network view of trafficking risk in commercial sex supply chains, from recruitment to sales.

Discussion

[3] We consider a business suspicious if it has contact information (website, phone number) that appears in the meta data of commercial sex sales advertisements in our deep web dataset.

We leverage machine learning and deep web data to construct the first large-scale and data-driven view of commercial sex supply chains. Our approach uniquely allows us to link deceptive recruitment activity to sex sales by the same entity to unmask trafficking risk. These results yield several policy-relevant insights.

First, inferring likely recruitment-to-sales pathways can help law enforcement agencies along potential trafficking routes better coordinate efforts. The FBI reports that the most effective way to investigate human trafficking is through a "collaborative, multi-agency approach with our federal, state, local, and tribal partners" [44]. For example, they hold an annual week-long counter-trafficking 'sweep,' where law enforcement officials across the United States respond undercover to sex sales posts to generate leads on traffickers. This synchronized effort has shown great success, leading to 67 arrests in 2019 [45], but it has its drawbacks. Naturally, a sustained counter-trafficking effort would be more effective; however, it is costly for many agencies to simultaneously collaborate in this fashion, and there is currently no systematic way to determine which collaborations to prioritize [12]. Also, sweeps are focused on major cities with high sex sales pressure, largely ignoring high-risk suburban locations with high recruitment pressure. Our analysis uncovers likely trafficking routes to help prioritize partnerships between impacted law enforcement jurisdictions; moreover, instead of focusing purely on sex sales, these collaborations can holistically tackle an entity's trafficking supply chain, from recruitment to sales.

Second, identifying region-specific exploitative behaviors can inform targeted local policies and interventions. Social policy plays an important role in preventing vulnerable victims from being trafficked [8], as well as rehabilitating victims after their trafficking experience [46]. While the latter (mitigation) is more prevalent, the former (prevention) shows significant promise since many victims are domestic, e.g., an estimated 67% of trafficking victims in the United States are United States citizens [47], and 93% of victims in Canada are Canadian citizens [48]. To this end, our results provide large-scale insight into where and how victims are (often deceptively) recruited. Cities with high recruitment pressure may prefer to focus their resources on preventative measures and can furthermore tailor interventions towards the recruitment tactics frequently seen in their specific locale. Prioritizing resource allocation to maximize impact in this manner is valuable since social resources are often highly constrained.

There are some limitations that may materialize if there is significant adoption of these methods in counter-trafficking. First, criminals may respond by creating new recruitment templates in order to evade detection. This can be combatted by periodically re-training the machine learning model using our active learning approach and ensuring up-to-date coverage of commercial sex websites. Second, sex trafficking entities may cease using the same contact information (i.e., meta data) across locations, making it more difficult to infer an organization's recruitment-to-sales pathways (although one can still reliably infer recruitment and sales pressure). In this case, new methods can be explored for mappings, e.g., based on shared post verbiage/style. We note that it is unlikely that criminals will respond with these shifts in the near term.

This work demonstrates how powerful machine learning tools can be applied in tandem with domain expertise for inference in settings with highly imbalanced and networked data. Our approach can be leveraged to investigate other type of trafficking with a heavy web presence (e.g., drugs, weapons, etc.) or, more broadly, in applications that require uncovering granular local patterns from large-scale, unstructured textual data.

Bibliography

[1] A. Walker-Rodriguez and R. Hill, "Human Sex Trafficking," FBI: Law Enforcement Bulletin, 2011.

[2] "Human trafficking by the numbers," International Labor Organization, 2017.

[3] "Human Trafficking: Better Data, Strategy, and Reporting Needed to Enhance U.S. Antitrafficking Efforts Abroad," *United States Government Accountability Office,* 2006.

[4] F. Laczko, "Human Trafficking: The Need for Better Data," *Migration Policy Institute,* 1 November 2002.

[5] M. De Witte, "The anti-trafficking movement needs better data to solve the problem, Stanford researchers say," Stanford News Service, 2018.

[6] C. Flynn, M. Alston and R. Mason, "Trafficking in women for sexual exploitation: Building Australian Knowledge," International Social Work, 2012.

[7] D. Androff, "The problem with contemporary slavery: An international human rights challenge for social work," International Social Work, 2010.

[8] J. Orme and F. Ross-Sheriff, "Sex trafficking: Policies, programs, and services," Social Work, 2015.

[9] K. Kotrla, "Domestic minor sex trafficking in the United States," Social Work, 2010.

[10] M. Potocky, "The travesty of human trafficking: A decade of failed U.S. policy," Social Work, 2010.

[11] J. Roby and M. Vincent, "Federal and state responses to domestic minor sex trafficking: the evolution of policy," Social Work, 2017.

[12] M. Shively, K. Kliorys, K. Wheeler and D. Hunt, "A National Overview of Prostitution and Sex Trafficking Demand Reduction Efforts, Final Report," The National Institute of Justice Office of Justice Programs, U.S. Department of Justice, 2012.

[13] L. Murphy, "Labor and Sex Trafficking Among Homeless Youth," Modern Slavery Research Project, Loyola University New Orleans, 2017.

[14] D. Baker and E. Grover, "Responding to victims of human trafficking: Interagency awareness, housing services, and spiritual care," Social Work & Christianity, 2013.

[15] T. Heilemann and J. Santhiveeran, "How do female adolescents cope and survive the hardships of prostitution? A content analysis of existing literature," Journal of Ethnic & Cultural Diversity in Social Work, 2011.

[16] D. Hodge and C. Lietz, "The international sex trafficking of women and children: a review of the literature," Affilia, 2007.

[17] B. Johnson, "Aftercare of survivors of human trafficking," vol. 39, pp. 370-389, 2012.

[18] L. Jones, D. Engstrom and T. Hilliard, "Globalization and human trafficking," *Journal of Sociology and Social Welfare,* vol. 34, no. 2, pp. 102-122, 2007.

[19] J. Roby, "Women and children in the global sex trade: toward a more effective policy," *International Social Work,* vol. 48, no. 2, pp. 136-147, 2005.

[20] D. Okech, Y. Joon Choi, J. Elkins and A. Burns, "Seventeen years of human trafficking research in social work: a review of the literature.," Journal of Evidence-Informed Social Work, 2018.

[21] S. Raets and J. Janssens, "Trafficking and technology: Exploring the role of digital communication technologies in the belgian human trafficking business," European Journal on Criminal Policy and Research, 2019.

[22] M. Latonero, "Human Trafficking Online The Role of Social Networking Sites and Online Classifieds," USC Annenberg Center on Communication Leadership & Policy, 2011.

[23] "Global Report on Trafficking in Persons 2020," United Nations Office on Drugs and Crime, 2020.

[24] E. Albright and K. D'Adamo, "Decreasing Human Trafficking through Sex Work Decriminalization," *AMA Journal of Ethics,* vol. 19, no. 1, pp. 122-126, 2017.

[25] M. Dank, B. Khan, M. Downey, C. Kotonias, D. Mayer, C. Owens, L. Pacifici and L. Yu, "Estimating the Size and Structure of the Underground Commercial Sex Economy in Eight Major US Cities," *The Urban Institute,* 2014.

[26] "TellFinder alliance: a global counter-human trafficking partner network, empowered by data," 2021. [Online]. Available: https://www.tellfinderalliance.com/.

[27] M. Kejriwal and R. Kapoor, "Network-theoretic information extraction quality assessment in the human trafficking domain," *Applied Network Science,* vol. 4, no. 44, 2019.

[28] C. Zhang, C. Re, M. Cafarella, C. De Sa, A. Ratner, J. Shin, F. Wang and S. Wu, "DeepDive: declarative knowledge base construction," *Communications of the ACM,* 2017.

[29] E. Hall, C. Dickson and D. Schroh, "TellFinder: Discovering related content in big data," *Vis,* 2015.

[30] Z. Zhou, "A brief introduction to weakly supervised learning," *National Science Review,* vol. 5, no. 1, pp. 44-53, 2018.

[31] A. Ratner and et. al., "Snorkel: fast training set generation for information extraction," *Proceedings of the 2017 ACM International Conference on Management of Data,* 2017.

[32] J. Gonsior, M. Thiele and W. Lehner, "WeakAL: Combining Active Learning and Weak Supervision," International Conference on Discovery Science, 2020.

[33] M. Nashaat and J. Miller, "Improving News Popularity Estimation via Weak Supervision and Metaactive Learning," Proceedings of the 54th Hawaii International Conference on System Sciences, 2021.

[34] T. Mikolov and et al., "Efficient estimation of word representations in vector space.," arXiv preprint arXiv:1301.3781, 2013.

[35] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* pp. 1532-1543, 2014.

[36] W. Yang, W. Lu and V. Zheng, "A simple regularization-based algorithm for learning cross-domain word embeddings.," arXiv preprint arXiv:1902.00184, 2019.

[37] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," Proceedings of the LREC 2010 Workshop on New Challenges for NLP, 2010.

[38] A. Ratner, S. Bach, H. Ehrenberg, J. Fries, S. Wu and C. Re, "Snorkel: rapid training data creation with weak supervision," Proceedings of the VLDB Endowment, 2017.

[39] M. Craven and J. Kumlien, "Constructing biological knowledge bases by extracting information from text sources," *ISMB,* pp. 77-86, 1999.

[40] B. Settles, "Active learning," *Synthesis lectures on Artificial Intelligence and Machine Learning,* vol. 6, no. 1, pp. 1-114, 2012.

[41] B. Settles, "Active Learning Literature Survey," Computer Sciences Technical Report: University of Wisconsin Madison, 2009.

[42] D. Otter, J. Medina and J. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems,* 2020.

[43] J. Zhu, H. Wang, E. Hovy and M. Ma, "Confidence-Based Stopping Criteria for Active Learning for Data Annotation," *ACM Transactions on Speech and Language Processing,* vol. 6, no. 3, 2010.

[44] "Human Trafficking," FBI.gov, [Online]. Available: https://www.fbi.gov/investigate/violent-crime/human-trafficking. [Accessed 03 2021].

[45] "Operation Independence Day," FBI.gov, 6 August 2019. [Online]. Available: https://www.fbi.gov/news/stories/operation-independence-day-2019.

[46] Y. Rafferty, "The Impact of Trafficking on Children: Psychologicaland Social Policy Perspectives," *Child Development Perspectives,* vol. 2, no. 1, pp. 13-18, 2015.

[47] S. Jorgensen and P. Sandoval, "Experts: Trump's tape bound women trafficking claim is misleading," CNN, 2019.

[48] M. Lopez-Martinez, "Sex trafficking still a prevalent issue a cross Canada, advocates and police say," CTV News, 2020.

Acknowledgments: The authors are grateful to Chris Dickson and Danielle Smalls of Uncharted Software for assistance with curating and contextualizing the core deep web dataset, as well as Carolina Holderness and Pierre Griffith of the Human Trafficking Response Unit at the Manhattan District Attorney's Office for providing invaluable domain insights. The authors also thank Tsai-Hsuan Chung for collecting auxiliary data in support of this work, and David Jonker for helpful comments on an earlier draft.

Funding:

Manhattan District Attorney's Office 20180100003 contract (EW)

Wharton Analytics Initiative and Social Impact Initiative (PR, HB)

Author contributions: PR: conceptualization, methodology, software, formal analysis, writing; HB: conceptualization, methodology, supervision, writing; EW: conceptualization, resources, writing, funding acquisition.

Competing interests: PR serves as a Responsible AI director at PricewaterhouseCoopers, and EW serves as director of counter-human trafficking initiatives at Uncharted Software and co-chair of the TellFinder Alliance. HB declares no competing interests.

Data and materials availability: An aggregated and de-identified version of our dataset (for reconstructing commercial sex supply chain networks and evaluating trafficking risk) and code (implementing our machine learning framework) is available in the supplementary materials.

Raw deep web data is available from Uncharted Software, but restrictions apply to the availability of this data since they contain personal identifiable information. This data will be made available by special request by contacting the manager of risk and compliance, Nicole Hankel (nhankel@uncharted.software).

Supplementary Materials for

Unmasking human trafficking risk in commercial sex supply chains with machine learning

This appendix is organized as follows. Section 1 describes our methods in detail, Section 2 provides details on our deep web dataset, and Section 3 presents auxiliary results.

1 Methods

Recall that our machine learning framework proceeds as follows: construction of an initial well-balanced training set via weak learners, initial predictive model selection, a custom active learning approach aimed at network discovery, and construction of supply chain networks through shared identifiers in the meta data. We now describe each step in detail.

Initial Training Set

Our entire dataset of 13,568,130 posts is initially unlabeled. Labels must be obtained by a domain expert manually reading the content of a post and assigning a label (recruitment vs. sales); recruitment posts are additionally tagged with the type of recruiting tactic (see Table S3 for a list of the 27 recruiting categories we ultimately identify). The sensitive nature of the data (i.e., containing personal identifiable information such as names and contact information) precludes a crowd-sourcing approach. To the best of our knowledge, there has been no prior work in academia or industry on predicting trafficking risk in recruitment using machine learning. As a result, we cannot apply existing models to our unlabeled data. Thus, we must obtain manual labels for a subset of our posts in order to obtain an initial training set to build a predictive model.

A common approach is to label a random subset of the posts to create this initial training set [1]. However, as noted in the main text, we estimate that over 99.4% of posts are aimed at sex *sales*; this is to be expected since our dataset is collected from leading commercial sex advertisement websites. Consequently, a domain expert would have to manually label nearly 2000 randomly chosen posts to find a *single* instance of recruitment with reasonable likelihood. More advanced sampling approaches – e.g., sampling from clusters of the data [2] or dense regions [3], or maximizing diversity in the sample [4] – experience similar imbalance issues in constructing an initial training set. This is problematic because, from a statistical perspective in a classification problem, the *effective sample size* of the data scales with the number of observations in the minority class (i.e., the number of labeled recruitment posts). At the same time, we leverage deep

learning models (due to their incomparable success on prediction with unstructured text [5]), which have a high tendency to overfit training data and are therefore data hungry [6].

Thus, we must carefully choose a subset of posts that (i) has a far higher likelihood of containing recruitment posts, thereby ensuring that our initial training set has a nontrivial effective sample size, and (ii) is of a manageable size for manual labeling by domain experts. To address this issue, we construct an initial 'recruitment vocabulary' that informs a human-in-the-loop weak learning approach.

First, we must preprocess the text to capture the semantic content of words in a way that can be passed as an input to a machine learning algorithm. A leading approach is to train *word embeddings*, which project words into a vector space whose distance metric captures semantic similarity [7]. Typically, word embeddings are trained to encode how frequently pairs of words co-occur in text; this is an effective approach since words with similar meanings tend to occur in similar contexts. To capture the unique context of our data, we train our own domain-specific word embeddings using Gensim word2vec [8, 9], which uses "continuous bag of words" (CBOW). Specifically, CBOW involves specifying a window of "context words" around a "target" word that are used to predict the target. Model parameters are iteratively updated using different pairs of context-target word combinations to modify a target word's embedding, based on its appearance with its co-occurring neighbors [10]. Following standard pre-processing techniques in natural language processing [11], we drop stop words (e.g., 'the', 'is', 'but') and lemmatize the vocabulary (e.g., 'caring' would be converted to care, 'communicating' would be converted to communicate). This leaves a unique vocabulary of size 223,883 across all posts. Using a context window size of 5, we train embeddings of dimension 100.

Then, we identify some candidate terminology that signifies recruitment risk from discussions with domain experts. These words were chosen through a human-in-the-loop process to maximize the likelihood of the corresponding post being recruitment-related; thus, words such as 'model' that are likely to appear in both recruitment and sales posts were excluded in order to avoid a high false positive rate. Our initial 'recruitment vocabulary' includes all terms whose embeddings are within a short distance of the embeddings corresponding to the expert-identified terms shown in Table S1.

Table S1: List of expert-identified recruitment terms used to inform weak learners

'audition', 'salary', 'interview', 'earn money', 'high pay', 'scout', 'staff', 'paid', 'employees', 'salaries', 'working', 'opportunity', 'earning', 'recruiting', 'recruitment', 'recruiter', 'hiring', 'hire', 'airfare', 'applicants', 'airfare travel', 'renumeration', 'commission'

As with traditional weak supervised learning [12], the presence of a term from our recruitment vocabulary provides a noisy signal that a post may be related to recruiting. Using the Snorkel package [13] , we train a weak supervision model that results in 1651 posts containing part of

this recruitment vocabulary. We then obtain labels for this small subset of posts, resulting in 369 recruitment-related posts. Note that this corresponds to 22% of the labels being positive, compared to only 0.06% of the labels being positive on a random subsample of our dataset.

Predictive Model Selection

Before proceeding to our active learning strategy, we must select a machine learning model for prediction. Deep neural networks (DNNs) have shown great success in text classification tasks [5], but there are a number of state-of-the-art approaches that may be promising. Thus, we train and evaluate 6 types of DNN models using our initial training data: 4 rely only on our data alone, while 2 additionally incorporate transfer learning from existing language models.

To improve the quality of our initial predictive model, we also augment our initial training data by adding structured noise to the labeled posts. We leverage a series of transformation functions that replace names, adjectives, and verbs with synonyms in order to generate a set of synthetic labeled posts; such an approach is helpful when the training set is small because it helps the predictive model avoid overfitting to irrelevant features (e.g., names) [14]. However, as we collect additional data through active learning, we discard the synthetic posts generated by data augmentation in model training.

We reserve a 20% random subsample of our initial training data as a validation set, on which we evaluate the predictive quality of all 6 models (see results in Table S2). The first four models are built using Keras in Tensorflow [15]. The base model ("Model 1") takes an input of tokenized sequences that represent each post. First, the input enters an embeddings layer that allows the model to modify the word-vectors used to encode the text during model training while it learns which posts are likely recruitment. The learned embeddings are then fed into a global average pooling layer to help prevent overfitting [16]. The final layer is a densely connected layer with a sigmoid activation function, which is useful for predicting probabilities [17]. Our second model ("Model 2") additionally includes dropout (a regularization method in which some number of nodes in the deep neural network are ignored during training), which has been shown to reduce overfitting to the training set [18]. We also include a bias initializer to help address the remaining data imbalance in our training set [19]. Next, we test two long short-term memory (LSTM) models (a type of recurrent neural network that is capable of learning the order dependence in a sequence), which are useful for text classification [20]. We test both a simple LSTM ("Model 3") [21], and a bi-directional LSTM ("Model 4") that leverages both the input sequence and a reversed copy in order to learn the whole context [22]. Finally, we test transfer learning from two state-of-the-art language models, BERT [23] and XLNET [24], using the Simple Transformers package [25]. Transfer learning allows us to take advantage of pre-training on larger datasets and fine-tune a model to our particular classification task [26]. The results of the 6 models tested are shown in Table S2.

Table S2: Comparison of models trained and evaluated on initial training dataset. Note that these values are based on our initial model, and our model improves as we collect data through the active learning process. The selected model is shown in bold.

Model	*Validation Precision*	*Validation Recall*	*Validation Accuracy*
Model 1	89.3%	79.3%	92%
Model 2	**91.2%**	**82%**	**93.7%**
Model 3	88.6%	80.2%	94%
Model 4	83.8%	80%	93%
BERT	55%	72%	86%
XLNET	67%	65%	89%

On an imbalanced dataset, one can achieve high accuracy by simply always predicting the majority class. Rather, our goal is to identify as many recruitment-related posts as possible. Therefore, a predictive model that has many false negatives (recruitment posts that are predicted to be sales posts) is especially undesirable. Thus, we select Model 2 – which has the highest precision and recall on the validation set of all the models we tested – to be our predictive model class to use in the active learning process.

Active Learning

As noted in the main text, our initial training dataset – and the resulting predictive model – is clearly biased by the purview of domain experts and does not provide a complete view of the numerous styles/tactics of recruiting posts on the deep web. Thus, we use pool-based active learning [27] to iteratively identify new, uncertain posts to label in order to improve the quality/coverage of our predictive model while maintaining a feasible labeling burden. Importantly, to achieve good coverage of the entire supply chain network across locations, we modify the prioritization function relative to traditional active learning.

We first define some notation. Let X be the pool of all posts; at any point of time, this pool is composed of mutually exclusive sets $X = X_0 \cup X_1$, where X_0 is the set of unlabeled posts and X_1 is the (much smaller) set of labeled posts with corresponding binary labels Y_1. Each post $x \in X$ is associated with two quantities: a (potentially empty) set of locations L_x and a (potentially empty) set of identifying information M_x (e.g., phone number, email). 94% of posts in our sample have at least one location and 69% of posts have at least one identifier.

Then, for every unlabeled post $x \in X_0$, we can construct the set of potential "edges" (i.e., between a pair of locations) that it may inform for network discovery. We are specifically interested in edges in the commercial sex supply chain network that may carry trafficking risk. Thus, we define the set:

$$E_x = \{\ell_1 \leftrightarrow \ell_2 \mid \exists x' \in X \ s.t. \ M_x \cap M_{x'} \neq \emptyset \ \ and \ \ell_1 \in L_x, \ell_2 \in L_{x'}\}.$$

In other words, for any unlabeled post $x \in X_0$, E_x captures the number of potential recruiting-sales or recruiting-recruiting location edges we will identify (based on some shared identifier from the meta data M_x) *if* x is found to be a recruiting post. Note that some of these edges may already be known to carry (or likely not carry) trafficking risk based on other labeled samples.

For each batch in active learning, we iteratively re-train our selected model (DNN with imbalance weights) using the currently labeled posts (X_1, Y_1) as our training set; this yields a model $f: x \to (0,1)$ that predicts the likelihood that a post x is a recruiting post based on its text.[1] Then, we apply the model to predict the probability $f(x)$ that each currently unlabeled post $x \in X_0$ is a recruiting post. Traditional active learning would solely rely on this metric to determine which posts to prioritize for labeling – specifically, we define the function

$$\chi(x) = 1 - \left|\frac{1}{2} - f(x)\right|.$$

$\chi(\cdot)$ captures the *uncertainty* of a post's prediction. Traditional active learning seeks to reduce labeling effort by focusing effort away from posts that already have confident predictions (e.g., clearly sales, $f(x) \approx 0$, or clearly recruitment, $f(x) \approx 1$) and are therefore unlikely to improve the accuracy of the current predictive model. Instead, active learning prioritizes posts $x \in X_0$ that have high values of $\chi(x)$ (i.e., values of $f(x)$ that are close to 1); these are the posts for which the current predictive model is relatively uninformative, and therefore augmenting the training set with the labels of these posts may improve the accuracy of the model.

However, such an approach focuses purely on improving the predictive accuracy of f across all posts. As noted earlier, our objective is more nuanced – we seek to uncover an accurate representation of the underlying supply chain network *across* locations. We address this objective mismatch by incorporating geographical diversity and the likelihood of identifying new network edges in our learning procedure. Unlike traditional active learning, our new prioritization will crucially rely on the meta data (L_x, M_x) associated with a post x.

Thus, in addition to $\chi(\cdot)$, we define additional metrics – the 'node uncertainty' and the 'edge uncertainty' to capture how a post contributes to geographically diverse coverage. To formalize these metrics, we require some additional notation. We begin by defining two useful subsets of unlabeled posts:

$$\Delta = \{x \in X_0 \mid 0.4 \leq f(x) \leq 0.8\},$$

$$\eta = \{x \in X_0 \mid f(x) > 0.8\}.$$

Δ captures uncertain posts, while η captures likely-recruitment posts. The upper and lower bounds are tuning parameters that were chosen to optimize the performance of the active learning procedure.

[1] Technically, f is trained on the word embedding of x, but we abuse notation for clarity.

Node Uncertainty: Let the set of unlabeled posts corresponding to a certain location be denoted by

$$V(\ell) = \{x \in X_0 \mid \ell \in L_x\},$$

and the set of unlabeled posts corresponding to a certain edge be denoted by

$$T(e) = \{x \in X_0 \mid e \in E_x\}.$$

Then for a given location ℓ, we define the 'node uncertainty' to be

$$N(\ell) = \frac{|\Delta \cap V(\ell)|^2}{|\eta \cap V(\ell)| + 1}.$$

$N(\ell)$ captures the extent to which we distinguish potential recruitment tactics at location ℓ. Specifically, if the numerator (number of uncertain posts associated with location ℓ) is high, we wish to prioritize posts associated with this location; in contrast, if the denominator is high (we have identified many likely-recruitment posts already), we wish to de-prioritize associated posts. Then, for every unlabeled post $x \in X_0$, we can compute a normalized score of how much labeling it may contribute to reducing node uncertainty for its set of locations L_x:

$$N(x) = \frac{1}{|L_x|} \sum_{\ell \in L_x} N(\ell).$$

Edge Uncertainty: Analogously, let the set of unlabeled posts corresponding to a certain pair of locations be denoted by

$$T(e) = \{x \in X_0 \mid e \in E_x\}.$$

Then, for a given edge e between a pair of locations, we define the 'edge uncertainty' to be

$$M(e) = \frac{|\Delta \cap T(e)|^2}{|\eta \cap T(e)| + 1}.$$

$M(e)$ captures the extent to which we distinguish potential recruitment-to-sales or recruitment-recruitment pathways in the supply chain network for an edge e between a pair of locations. If the numerator (number of uncertain posts associated with edge e) is high, we wish to prioritize posts associated with this location; in contrast, if the denominator is high (we have identified many likely-recruitment posts already), we wish to de-prioritize associated posts. Then, for every unlabeled post $x \in X_0$, we can compute a normalized score of how much labeling it may contribute to reducing edge uncertainty for its set of locations E_x:

$$M(x) = \frac{1}{|E_x|} \sum_{e \in E_x} M(e).$$

Active Learning: Our active learning strategy proceeds in batches. In each batch, we use the current predictive model f to make predictions on every currently unlabeled post $x \in X_0$. All

likely-recruitment posts are automatically prioritized for labeling. Following traditional active learning, we also prioritize posts with high prediction uncertainty $\chi(x)$. Then, to improve network discovery, we prioritize posts that have a high score $N(x)$ for reducing node uncertainty, and a high score $M(x)$ for reducing edge certainty in our supply chain network. We rank the unlabeled posts in $x \in X_0 \cap \eta^c$ according to our modified metrics and choose the top ~4,000 posts to label. Once these labels are obtained, we appropriately modify X_0, X_1 and retrain our deep learning model f on the augmented training set X_1. We then re-compute our set of unlabeled likely-recruitment posts η; we stop the active learning process when this set is empty.

Active Learning Pseudocode

Input: *unlabeled posts* X_0*, labeled posts* X_1*, initial model* f *trained on initial training set*

Predict $f(x)$ for every $x \in X_0$
Compute the set of unlabeled likely-recruitment posts η
while $\eta \neq \emptyset$ **do**
    Initialize prioritized posts for labeling to $B = \eta$
    Compute 'prediction uncertainty' $\chi(x)$ for every remaining $x \in X_0 \cap \eta^c$
    Compute 'node uncertainty' $N(\ell)$ for every location ℓ
    Compute 'edge uncertainty' $E(e)$ for every edge e
    Compute $N(x), M(x)$ for every remaining $x \in X_0 \cap \eta^c$
    Sort remaining posts $x \in X_0 \cap \eta^c$ by descending order of $\chi(x)$ and then by descending order of $N(x), M(x)$
    Select top ~4000 posts P and add to batch to be labeled $B \leftarrow B \cup P$
    Obtain manual labels (x, y) for all $x \in B$
    Update labeled set $X_1 \leftarrow X_1 \cup B$, and unlabeled set $X_0 \leftarrow X_0 \cap B^c$
    Train new predictive model $f(\cdot)$ using augmented training data (X_1, Y_1)
    Compute the set of unlabeled likely-recruitment posts η
end

We ran 13 batches of active learning.[2] Figure S1 shows the histogram of prediction scores $f(x)$ on our unlabeled posts X_0 after each batch. In the first batch (after training on only our initial training set), we observe a very large spread of prediction scores across the interval $(0,1)$, indicating a large degree of uncertainty. In later batches, as we iteratively label both likely-recruitment and uncertain posts, we observe that the number of likely-recruitment posts (i.e., predictions above 0.8) among the unlabeled set X_0 decreases steeply as the results converge (note the scale of the y-axis across batches in Figure S1). Figure S2 shows how the accuracy of f evolves with each additional batch; note that its performance is asymptotes, suggesting convergence.

[2] In earlier batches, if B was very large, we only obtained labels on a random subset of B.

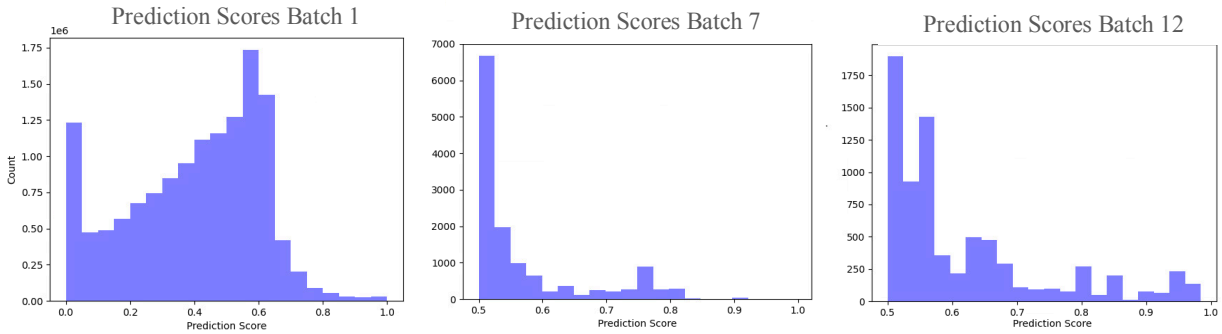

Figure S1: Prediction scores from our model on the unlabeled data for three batches across the process. Note that the y-axis and x-axis are different across plots. Batch 12 has far fewer uncertain posts (prediction score above .5) than any prior batch.

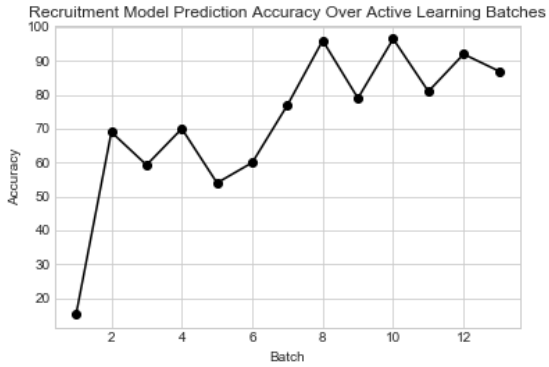

Figure S2: Accuracy of the predictive model f across 13 batches of the active learning process.

Throughout the process, we obtained labels on a total of 50,199 total posts, of which 6,953 posts were identified as recruitment. This corresponds to 14% of the labels being positive, as opposed to <0.1% if we had labeled randomly. Critically, while our initial training set only identified 3 types of recruitment tactics, our active learning process uncovered 27 recruitment tactics (see Table S3), demonstrating the effectiveness of our proposed approach. We additionally note that relying on traditional active learning alone would have directed our labeling efforts to posts from large cities (where the majority of posts occur), missing out on key recruitment tactics we identify in smaller cities (where we actually find recruitment dominates).

Table S3: List of recruitment tactics identified across all labeled posts.

Category	Definition
Adult Entertainment	Entertainment companies, bars, restaurants, strip clubs, bachelor parties, etc.
Escort	Agencies identified as escort services
Personal	Ads posted by individuals requesting personal interactions
Modeling	Agencies specifying jobs related to modeling

Porn	Ads recruiting for filming pornography
Massage	Ads recruiting for spas or massage parlors
Sugar	Ads recruiting for a sugar baby, a relationship where an individual provides money in exchange for an on-going relationship
Non-specified agency	Ads recruiting without specifying the type of work or job
Housing	Ads recruiting for vacant housing
Promotions	Job related to promoting products
Product Advertisement	Recruitment related to advertising products
Companionship	Ads specifying a paid companionship
House-keeping	Recruitment for house cleaning or cooking
Partnership	Ads recruiting for a business partner or escort partner
Make money	Ads specifying they can help you make money quickly
Walking	Recruitment for getting paid to walk
Booker	Recruitment for being a booker for an agency
Photography	Recruitment for exchanging photography for services
New Venture	Ads specifying partnering on a new venture
Finance	Recruitment for finance jobs
Club	Recruitment to join a specific club
Gangbang	Recruitment to be paid for a gang bang
Corporate Fitness	Corporate fitness jobs
Asian job	Roles specifying recruiting Asian women
Tourism	Recruitment for jobs related to hotels or tourism
Contest	Recruitment for contests
Videochat	Recruitment to get paid for a videochat

Network Creation

Finally, we connect the recruitment posts we identified with the remaining posts (which we know are over 99.9% sales) to uncover likely recruitment-to-sales pathways in commercial sex supply chains. We use the following variables extracted from the meta data of posts to identify entities: email, phone number, username, URL, and social media handle. We find 43,521 connections in total from recruitment to commercial sex sales posts; surprisingly, 10% of recruitment posts account for 85% of the connections.

2 Deep Web Data

Our core deep web dataset is obtained from our collaborators at the TellFinder Alliance for global counter-human trafficking. The deep web consists of (often temporary) pages that are not indexed by Google, and therefore need to be scraped in real-time. TellFinder works with its

partners in law enforcement to identify websites with significant commercial sex activity – which often carry risk of exploitation and human trafficking [28] – that are relevant to counter-trafficking efforts. They leverage recent technology developed to scrape deep web data, extract relevant meta data (e.g., phone numbers, email addresses) and convert it into databases that support trafficking investigation inquiries by law enforcement agencies [29, 30].

There are several kinds of websites where commercial sex activity can be deduced. These include service review websites such as the Erotic Review or Rubmaps and discussion forms (where content is largely shared by *consumers*, rating specific sexual services), as well as commercial sex advertisement websites (where content is largely shared by *entities* selling sexual services). Since our primary goal is to identify *supply chains* of specific entities – i.e., connect deceptive recruitment offers to sex sales by the same entity in order to pinpoint human trafficking risk – we focus on commercial sex advertisement websites where we can extract identifying information (e.g., phone numbers, emails) of the entities selling sex. Table S4 shows summary statistics of different identifiers extracted from posts on our deep web dataset; indeed, we see that the large majority of posts contain identifying information that can be used to connect entity-specific activity.

Table S4: Deep web data sample summary

Identifier	# Unique Occurrences	# Posts including this Identifier	% Posts including this Identifier
Phone number	393,132	8,503,617	62.7%
Email	214,728	1,489,803	11.0%
Social Media Handle	8,645	395,547	2.92%
Username	44,007	44,235	0.33%
Location	1,364	12,716,641	93.7%

One may not a priori expect that significant *recruitment* activity occurs on websites that primarily advertise commercial sex. We learned of this behavior from our law enforcement partners (e.g., when running a phone number associated with a criminal case through the TellFinder tool, one partner found that the number was associated with both sales posts and deceptive recruitment offers on the same website, thereby providing supporting evidence that this was likely a human trafficking case). However, as discussed in the main text, it was not possible to study these recruitment offers at scale since they are extremely rare; our machine learning framework addresses this challenge, and indeed identifies thousands of deceptive recruitment offers spanning many different tactics on these websites. These included job postings, personal ads, or ads offering other types of skills (see Table S3). Separately, we also examined posts on Craigslist.com, SpaStaff.com, and Indeed.com, where general recruitment activity is common (e.g., in the jobs and services categories) and content is not focused on commercial sex sales. If identifiers extracted from commercial sex advertisement websites match identifiers for entities recruiting on the websites for non-sex employment, this can also suggest

high risk for sex trafficking. However, we found only 4 such matches to Craigslist and no matches to SpaStaff.com or Indeed.com, suggesting that this behavior is relatively uncommon.

Our deep web dataset spans four websites that advertise commercial sex through English language posts. These posts were collected over a 9-month period spanning July 1, 2017 to September 6, 2018. The websites, in order of volume, include:

- www.skipthegames.com
- www.cityxguide.app
- www.megapersonals.eu
- www.adultwork.com

The resulting dataset comprises of 13,568,130 posts over 428 unique days. Figure S3 shows the breakdown of posts across website and location.

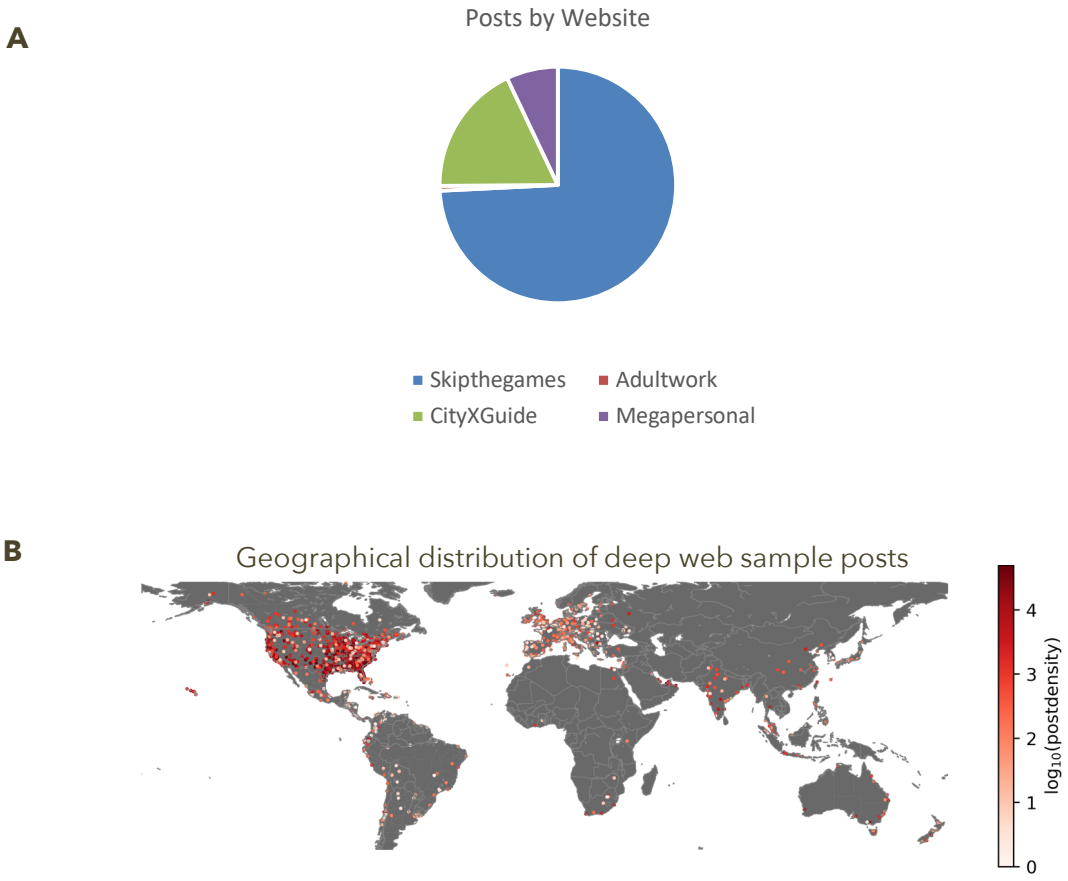

Figure S3: Count of deep web posts in our dataset by (A) website, and (B) location.

Figure S3 shows a key limitation of focusing on English language posts: the geographical distribution of our data is concentrated in countries with large English-speaking populations. In particular, approximately 95% of the posts are from the United States, Canada, the United Kingdom and Australia. We do find significant sales and recruitment activity in the rest of

Europe and in India, but this may be a biased sample since it omits activity occurring in local languages in those countries. A promising direction of future work is adapting our approach to other languages to improve global coverage. Note that this would require domain experts who speak these local languages to operationalize our human-in-the-loop and active learning steps.

3 Auxiliary Results

We now present some auxiliary results that support the conclusions in our main text. First, we examine the socio-economic characteristics of sender and receiver cities in the United States. Second, we compare the recruitment/sales densities we inferred from the deep web against Rubmaps.ch (a popular review site for massage parlors with sexual services) as well as suspicious businesses identified through Google Places.

Sender vs. Receiver Cities

We examine a number of relevant socioeconomic indicators (summarized in Table S5) to understand the characteristics of locations (in the United States) where vulnerable populations are deceptively recruited vs. sold for commercial sex. This data was collected at the county- or city-level across 8 government sources: US Census [31], US Bureau of Economic Analysis [32], US Bureau of Labor Statistics [33], US Department of Housing and Urban Development [34], National Center for Education Statistics [35], WomensShelters.org [36], Proximity One [37], and US Department of Justice [38]. The data collected from these sources focuses on both economic attributes (household income, GDP, unemployment) and social attributes (homelessness, education, crime).

Table S5: List of variables from government sources to compare the attributes of top recruitment cities against top sales cities in the United States. We employ the Benjamini Hochberg procedure to correct for multiple hypothesis testing.

Variable	Source	Kolmogorov Smirnov p-value	Statistical Significance after Benjamini Hochberg
Population	US Census (2018)	***	Yes
Real GDP	US Bureau of Economic Analysis (2018)	**	Yes
% of population with private health insurance	Proximity One (2009)	**	Yes
% of population with no health insurance	Proximity One (2009)	**	Yes
Violent crimes per 1000 people	US Department of Justice (2016)	**	Yes

Property crimes per 1000 people	US Department of Justice (2016)	**	Yes
Median household income	US Census (2018)	*	Yes
Poverty percent	US Census (2018)	*	Yes
Homeless per 1000 people	US Department of Housing and Urban Development (2019)	*	Yes
Homeless under 18 years old per 1000 people	US Department of Housing and Urban Development (2019)	*	Yes
Sheltered homeless per 1000 people	US Department of Housing and Urban Development (2019)	*	Yes
International migration per 1000 people	US Census (2018)	*	Yes
% of adults with bachelors degree	US Census (2018)		No
% of adults with less than high school education	US Census (2018)		No
% of adults with high school education	US Census (2018)		No
% of students granted Pell Grants (federal subsidy for college)	National Center for Education Statistics (2017)		No
Women's shelters per 1000 people	WomensShelters.org		No
Unemployment rate	US Bureau of Labor Statistics (2018)		No

p-value: *** <0.01, ** < 0.05, *<0.1

We run separate Kolmogorov Smirnov tests [39] to determine if there are systematic differences in the empirical distributions of each socioeconomic indicator in the top 50 'sender' versus top 50 'receiver' cities. We note that this is *not* a causal analysis since we are examining correlations. However, understanding the differences between recruitment and sales hubs can shed light on where different policy and social work interventions (e.g., those aimed at preventing victim recruitment vs. those aimed at rescuing current victims) would be the most impactful. Since we are testing a family of multiple related hypotheses, we employ the well-known Benjamini Hochberg procedure [40] to maintain the resulting false discovery rate (FDR) at a standard choice of 10%.

We find that sender cities tend to be smaller (lower populations) and economically more constrained (higher poverty and lower household incomes). Sender cities also have more homeless people (i.e., vulnerable populations) and suffer high crime incidence (both property crimes and violent crimes). Figure S4 highlights significant differences in variables amongst sender and receiver cities. Together, these results suggest sender cities may not have as many resources as larger receiver cities to prevent trafficking of their vulnerable populations. Thus, operationalizing collaborations between counter-trafficking agencies along inferred recruitment-to-sales trafficking routes may significantly benefit resource-constrained sender cities in preventing victims from being trafficked in the first place.

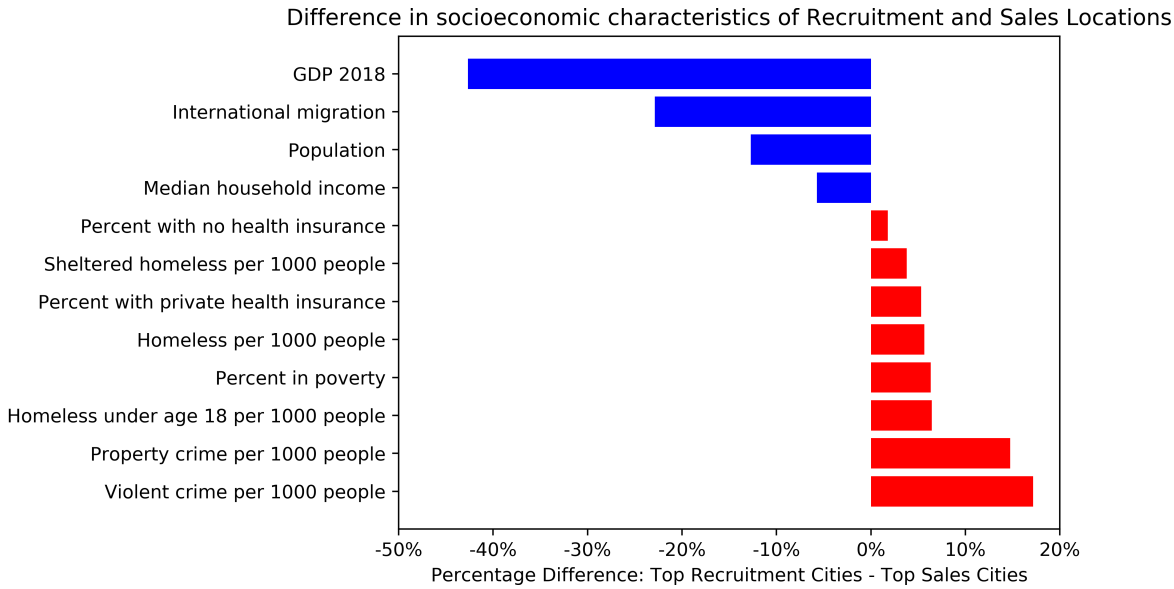

Figure S4: Comparing selected socioeconomic variables for the top 50 sender (recruitment) and receiver (sales) cities in the United States. Blue and red bars indicate variables with higher values in receiver and sender cities respectively.

Comparison to Rubmaps and Google Places

To the best of our knowledge, our study is the first to characterize recruitment in commercial sex supply chains, allowing us to uniquely identify trafficking recruitment risk at scale in commercial sex supply chains. In contrast, other empirical studies examine commercial sex activity purely from the sales side (e.g., through review websites such as Rubmaps), where the connection to human trafficking risk may be tenuous. We now examine how our deep web recruitment/sales densities compare to two such sources.

1. **Rubmaps.ch** is a review site for massage parlors with sexual services, and has been used to assess commercial sex activity in prior work [41, 42]. Rubmaps allows users to find and rate massage parlors by city/town. We manually extracted the count of massage parlors for each town listed on the website within the United States.

2. **Google Places**: Google Places includes a list of over 200 million global points of interest (e.g., restaurants, hotels, nail salons) that appear on Google Maps. We seek formally listed businesses with contact information (phone numbers or website) that also appear in the meta data of posts in our deep web dataset from commercial sex advertisement websites; in other words, these businesses are likely associated with commercial sex sales, and therefore we refer to them as suspicious businesses. We find 5035 suspicious businesses, with 2630 listed in the United States/Canada. We manually categorize these suspicious businesses and find that the majority are spa/massage parlors (55%); other significant categories include home services (e.g., cleaning, repair, pool, roofing, moving), dollar general stores, and law firms.

We map these datasets based on city names to obtain heat maps of commercial sex activity (see Figure S5). Of the top 50 receiver locations we identified in the United States using deep web data, 82% included locations of suspicious businesses found in Google Places and 46% included massage parlors identified in Rubmaps; in contrast, of the top 50 sender locations, only 72% overlapped with suspicious businesses in Google Places and 26% with Rubmaps. Thus, we find that commercial sex *sales* activity identified on the deep web roughly aligns with activity identified through Rubmaps and suspicious formal businesses that may be selling commercial sex; however, *recruitment* activity is distinct and uniquely identified by our analysis.

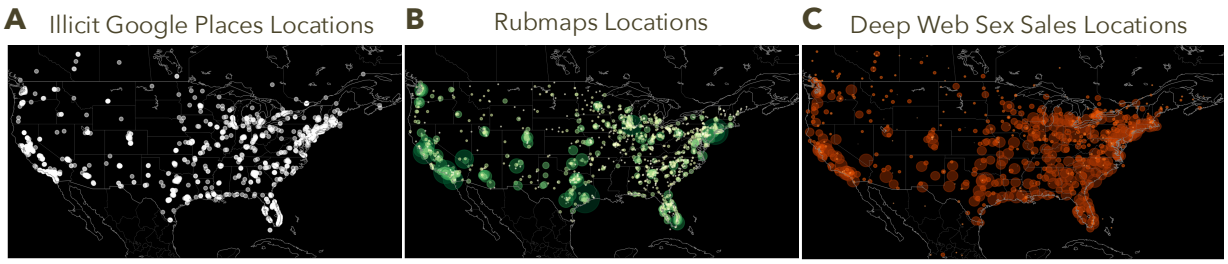

Figure S5: Empirical distribution of commercial sex activity in the United States inferred from (A) Google Places, (B) Rubmaps, and (C) Deep Web respectively.

Supplemental Materials Bibliography

[1] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," *Technical Report: Swedish Institute of Computer Science,* 2009.

[2] D. Dligach and M. Palmer, "Good seed makes a good crop: accelerating active learning using language modeling.," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* 2011.

[3] A. K. McCallum and K. Nigamy, "Employing EM and pool-based active learning for text classification," *Proc. International Conference on Machine Learning (ICML),* 1998.

[4] Y. Yang, Z. Ma, F. Nie, X. Chang and A. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision ,* 2015.

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu and J. Gao, "Deep Learning--based Text Classification: A Comprehensive Review.," *ACM Computing Surveys (CSUR),* vol. 54, no. 3, pp. 1-40, 2021.

[6] X.-W. Chen, "Big Data Deep Learning: Challenges and Perspectives," *IEEE Access,* vol. 2, pp. 514-525, 2014.

[7] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[8] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* pp. 45-50, 2010.

[9] "Gensim 4.0.1," [Online]. Available: https://pypi.org/project/gensim/.

[10] X. Rong, "word2vec parameter learning explained," *eprint arXiv:1411.2738,* 2014.

[11] S. Symeonidis, D. Effrosynidis and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," *Expert Systems with Applications ,* vol. 110, pp. 298-310, 2018.

[12] A. Ratner, S. Bach, H. Ehrenberg, J. Fries, S. Wu and C. Ré, "Snorkel: Rapid training data creation with weak supervision.," *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases,* vol. 11, no. 3, p. 269, 2017.

[13] "Snorkel," [Online]. Available: https://www.snorkel.org/.

[14] A. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon and C. Ré, "Learning to compose domain-specific transformations for data augmentation," *Advances in neural information processing systems,* vol. 30, p. 3239, 2017.

[15] "TensorFlow," [Online]. Available: https://www.tensorflow.org/.

[16] M. Lin, C. Qiang and Y. Shuicheng, "Network in network," *arXiv eprint arXiv:1312.4400,* 2013.

[17] C. Nwankpa, W. Ijomah, A. Gachagan and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning.," *arXiv eprint arXiv,* 2018.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research,* pp. 1929-1958, 2014.

[19] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence,* vol. 5, pp. 221-232, 2016.

[20] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification.," *Neurocomputing,* vol. 337, pp. 325-338, 2019.

[21] J. Schmidhuber and S. Hochreiter, "Long short-term memory.," *Neural Comput ,* vol. 8, pp. 1735-1780, 1997.

[22] M. Schuster and K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing ,* vol. 45, no. 11, 1997.

[23] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805 ,* 2018.

[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237,* 2019.

[25] "Simple Transformers," [Online]. Available: https://simpletransformers.ai/.

[26] C. Do and A. Y. Ng, "Transfer learning for text classification," *Advances in Neural Information Processing Systems,* vol. 18, pp. 299-306, 2005.

[27] A. McCallum and K. Nigamy, "Employing EM and pool-based active learning for text classification," *Proc. International Conference on Machine Learning (ICML),* pp. 359-367, 1998.

[28] "Global Report on Trafficking in Persons 2020," United Nations Office on Drugs and Crime, 2020.

[29] "TellFinder Alliance: a global counter-human trafficking partner network, empowerd by data," 2021. [Online]. Available: www.tellfinderalliance.com.

[30] E. Hall, C. Dickson, D. Schroh and W. Wright, "TellFinder: Discovering Related Content in Big Data," *VIS,* 2015.

[31] "U.S. Census," [Online]. Available: https://data.census.gov/cedsci/.

[32] "U.S. Bureau of Economic Analysis," [Online]. Available: https://apps.bea.gov/regional/downloadzip.cfm.

[33] "U.S. Bureau of Labor Statistics," [Online]. Available: https://www.bls.gov/lau/#tables.

[34] "US Department of Housing and Urban Development," [Online]. Available: https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/.

[35] "National Center for Education Statistics," [Online]. Available: https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7.

[36] "Women's Shelters," [Online]. Available: https://www.womenshelters.org/#state_list.

[37] "Proximity," [Online]. Available: http://proximityone.com/metro_healthinsurance.htm.

[38] "U.S. Department of Justice," [Online]. Available: https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-8/table-8.xls/view.

[39] N. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bulletin Moscow University,* vol. 2, no. 2, pp. 3-16, 1939.

[40] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological),* vol. 57, no. 1, pp. 289-300, 1995.

[41] V. Bouche and S. Crotty, "Estimating demand for illicit massage businesses in Houston, Texas," *Journal of human trafficking,* vol. 4, no. 4, 2018.

[42] M. Diaz and A. Panangadan, "Natural Language-based Integration of Online Review Datasets for Identification of Sex Trafficking Businesses," *IEEE 21st International Conference on Information Reuse and Integration for Data Science,* 2020.