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

(Authors' names blinded for peer review)

Problem Definition: 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. Existing research has focused on analyzing commercial sex sales on the internet to capture scalable geographical proxies for trafficking. However, ads selling commercial sex do not reveal information about worker consent. Therefore, it is challenging to identify risk for trafficking, which involves fraud, coercion or abuse.

Methodology: We leverage massive deep web data (collected globally from leading commercial sex websites) in tandem with a novel machine learning framework (combining natural language processing, active learning and network analysis) to study how and where sex worker *recruitment* occurs. This allows us to unmask potentially *deceptive* recruitment patterns (e.g., an entity that recruits for modeling, but sells sex), which signal high trafficking risk. We demonstrate via simulations that our approach outperforms existing active learning techniques to identify key nodes and edges in the underlying trafficking network. Our analysis provides a geographical network view of online commercial sex supply chains, highlighting deceptive recruitment-to-sales pathways that are likely trafficking routes.

Managerial Implications: Our results can help law enforcement agencies along trafficking routes better coordinate efforts to tackle trafficking entities at both ends of the supply chain, as well as target local social policies and interventions towards exploitative recruitment behavior frequently exhibited in that region.

Key words: human trafficking, machine learning, deep web, active learning, networks, text analytics

1. Introduction

The International Labor Organization estimates there were 4.8 million sex trafficking victims in 2017 alone (ILO 2017). Consequently, there is high demand from field experts (Office 2006, Laczko 2002, Witte 2018) and academics (Flynn et al. 2014, Androff 2011, Orme and Ross-Sheriff 2015, Kotrla 2010, Potocky 2010) for a large-scale and data-driven view of the underlying supply chain dynamics (Roby and Vincent 2017) 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) (Shively et al. 2012, Murphy 2016); this is in contrast to prevalent *mitigation* strategies that target the end of the supply chain (sales), e.g., coordinated stings on sex sales (FBI 2019), or the recent shutdown of the platform Backpage, which was heavily used to advertise sex sales (Kessler 2018). Furthermore, connecting recruitment activity to sex sales conducted by the same entity allows us to infer likely recruitment-to-sales *trafficking routes* of criminal entities; this can allow for powerful coordination strategies between relevant law enforcement agencies and task forces to increase the efficiency of counter-trafficking efforts by targeting both ends of the trafficking supply chain (Heilemann and Santhiveeran 2011, Hodge and Lietz 2007, Johnson 2012, Jones et al. 2007, Roby 2005).

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 (Okech et al. 2018). 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 (Androff 2011).

In this paper, we use unstructured, massive deep web data to characterize 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 (Raets and Janssens 2019, Latonero 2011), making the deep web a rich and relevant data source. Trafficking is commonly targeted at vulnerable populations (e.g., a study found that 1 out of 5 homeless youths in a North American homeless youth organization identified as victims of human trafficking (Murphy 2016)). These individuals are frequently recruited online through "fishing" strategies that offer well-paid jobs to attract potential victims to make initial contact with traffickers (Kangaspunta et al. 2020).

We begin by leveraging data from leading commercial sex advertisement websites in conjunction with a novel machine learning framework to construct a geographical network view of commercial sex *supply chains*, from recruitment to sales. Existing research has focused solely on analyzing commercial sex sales to capture proxies for trafficking (Dubrawski et al. 2015, Zhu et al. 2019). Importantly, however, commercial sex and sex trafficking are not synonymous (Albright and D'Adamo 2017): "Unlawful commercial sex acts overlap with sex trafficking when participation occurs by means of force, fraud, or coercion. . . " (Dank et al. 2014). 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 these challenges, we study how and where *deceptive recruitment* occurs to understand trafficking risk in commercial sex supply chains. Indeed, in over 50% of trafficking cases involving the internet analyzed by the UN Office of Drugs and Crime (UNODC) in 2020, victims reported making initial contact with a trafficker in response to a deceptive job advertisement.[1] For example,

[1] This report covers both sex and labor trafficking, but the majority of the examples included are for sex trafficking.

in one case, traffickers recruited approximately 100 women through a modeling job posting and then sex trafficked these women (Kangaspunta et al. 2020). Traffickers have also recruited for other adult services (e.g., stripping) before forcing victims into sex sales (UNODC 2020). Thus, 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 informative indicator that trafficking may have occurred. This proxy was informed through close collaboration with domain experts from the Tellfinder Alliance, ranging from members of human trafficking taskforces to policymakers. However, it is important to note that not all instances identified in this manner necessarily correspond to human trafficking; rather this is a high-quality proxy. Our work captures trafficking *risk* by linking likely deceptive (non-sex) recruitment offers to commercial sex sales by the same entities.

However, identifying recruitment content in ads has historically been a significant hurdle due to the nature of sex trafficking: while sex sales ads are prevalent, consistent in style, and convey clear intent to consumers, recruitment ads are sparse, vary significantly in style, 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 (TellFinder 2021, Kejriwal and Kapoor 2019, Zhang et al. 2017), 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, network analysis, and domain expertise to distinguish recruitment and sales content at scale. We then link deceptive recruitment activity to sex sales by the same entity to uncover their likely trafficking network. Notably, we introduce a new active learning strategy that is augmented by network metrics; we demonstrate through simulations that our approach performs considerably better than existing algorithms in identifying high-risk nodes (recruitment locations) and edges (recruitment-to-sales pathways) in the underlying trafficking network.

Our results on deep web data 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 (Murphy 2016). 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 help inform coordination strategies between relevant law enforcement agencies.

1.1. Additional Related Literature

The field of operations management is uniquely positioned to help tackle challenges in countering sex trafficking (Konrad et al. 2020). As described above, we take a supply chain perspective, aiming to understand how sex workers are recruited (supply) and sold (demand) by trafficking entities. A few studies have applied operations techniques to empirically and theoretically analyze other challenges in sex trafficking (Konrad et al. 2017). One important problem is effectively allocating resources for social policy interventions; Kaya et al. (2022) and Chan et al. (2018) use a multi-dimensional knapsack algorithm to improve access to housing and support services for homeless youth, with the goal of mitigating their vulnerability to trafficking. Maass et al. (2020) optimizes the placement of shelters for human trafficking survivors to maximize societal welfare while respecting budgetary constraints. Other work targets supporting law enforcement. For example, Li et al. (2021) uses online customer review data to support law enforcement in identifying massage businesses that are likely selling sex. Keskin et al. (2021) examines and predicts the movement patterns of entities selling commercial sex; this is an important goal to ensure that law enforcement officials do not pursue “wasted” stings on entities that have already left a location. Rabbany et al. (2018) use human-in-the-loop active learning to help law enforcement build trafficking cases, by iteratively connecting existing leads to other leads from a criminal database; in this case, the network is known, but the weights of each connection are unknown and must be learned by interacting with the domain expert. Kosmas et al. (2022) model interdiction strategies that can disrupt these illicit networks. Unlike past work which has focused on sex sales, our work identifies deceptive recruitment of victims, thereby providing an informative signal for distinguishing commercial sex and human trafficking.

The social sciences literature has also studied recruitment into sex trafficking, primarily via qualitative interviews with professionals, survivors of trafficking, legal cases, and online advertisements. For example, Baird and Connolly (2023) conducted a systematic review on sex trafficking recruitment of minors, and found that the most commonly cited recruitment location is the internet, but also identified a number of in-person recruitment locations (e.g., bus stops, homeless shelters, schools). Martin et al. (2014) noted that sex sales occur in very distinct places (e.g., hotels, cars, or abandoned buildings) compared to in-person recruitment locations. Gezinski and Gonzalez-Pons (2022) conducted a systematic review of online recruitment, and found sparse and noisy empirical evidence on the prevalence of sex trafficking recruitment online and noted that very few studies included specific references to how online recruitment occurred (e.g., website used, recruitment tactics, etc.). Our work complements this existing literature by providing large-scale insights into online deceptive recruitment tactics using massive deep web data.

Identifier	# Unique Occurrences	# Posts Including Identifier	% Post Including 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%

Table 1 Deep web data sample summary.

2. Deep Web Data

Our core deep web dataset is obtained from our collaborators at the TellFinder Alliance for global counter-human trafficking (TellFinder 2021). 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 (Kangaspunta et al. 2020) – 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 (TellFinder 2021, Hall et al. 2015).

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 1 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.

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, 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 likely deceptive recruitment offers spanning many different job categories on these websites. These included job postings, personal ads, or ads offering other types of skills (see Appendix D).[2]

Our deep web dataset spans four websites that advertise commercial sex through English language posts. These posts were collected over a 14-month period spanning July 1, 2017 to September 6, 2018. During this time period, on April 11, 2018, the United States government passed a law called FOSTA-SESTA that made it illegal to knowingly assist, facilitate or profit off of sex-trafficking. Some websites were completely removed (e.g. Backpage.com), while others made changes to the types of activity hosted (e.g. Craigslist removed personal ads). However, this did not stop online commercial sex activity; rather, advertising for the sex trade shifted to new websites (see, e.g., Kessler 2018), which made it even more important to provide updated trafficking risk models to law enforcement partners. We worked with the TellFinder Alliance to identify the top sites most frequently used *after* FOSTA-SESTA was enacted to ensure that our approach and results would be useful going forward. Indeed, in our data, 89% of the activity observed on these websites occurred after FOSTA-SESTA (reflecting increased activity), and we find that the network uncovered across the entire time period is representative of the post FOSTA-SESTA time period (see Appendix A for details). 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 1 shows the breakdown of posts across website and location.

Figure 1 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.

[2] Separately, we also examined posts on Craigslist.com, SpaStaff.com, and Indeed.com, with high recruitment activity (e.g., in the jobs and services categories) but 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.

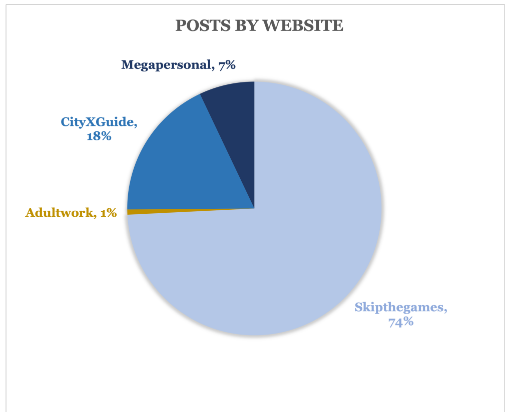


Figure 1 **Count of deep web posts in our dataset by website (top), and location (bottom).**

3. Machine Learning Framework

Our primary goals are to infer the underlying commercial sex supply chain and uncover varying types of recruitment from our deep web data sample. 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 advertise a job opportunity while sales posts advertise sexual services (see Figure 10 in §5 for examples of recruitment and sales posts). Recruitment posts are additionally manually categorized into types (e.g., sugar parent) based on the type of employment offer made. Manually labeling all 14 million posts is clearly infeasible; instead, we design an active learning approach to train a model with as few labels as possible. We face two challenges:

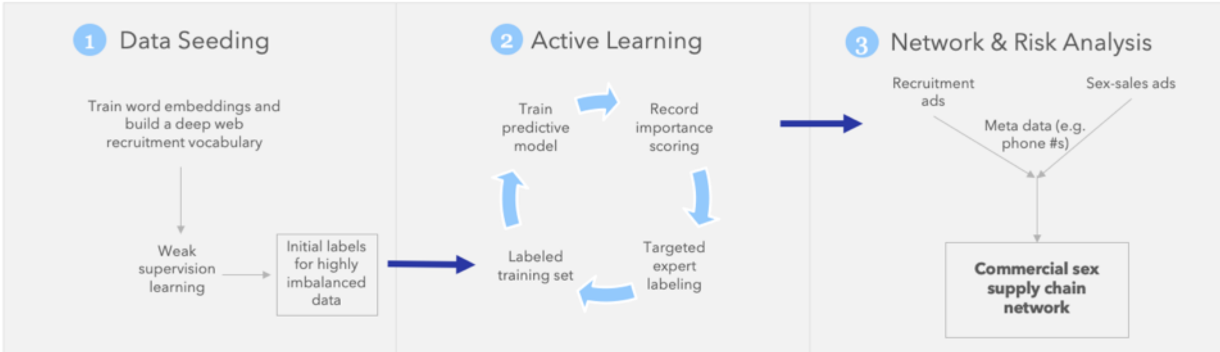


Figure 2 **Summary of our machine learning framework.We first train domain-specific word embeddings and collect expert-identified terms to develop a 'recruitment vocabulary.' This informs a weak learning heuristic to identify an initial well-balanced training set. We then apply active learning techniques (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.**

1. **Extreme Data Imbalance:** We estimate that somewhere between [0.05, 0.3]% of posts are recruitment-related (with 95% probability), while the rest are sales, i.e., one would have to manually label thousands of randomly chosen posts to find a handful of recruitment instances in expectation.[3] Thus, traditional supervised or active learning techniques, which rely on an initial well-balanced training set, are infeasible.

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.

We leverage weak learners (Zhou 2018, Ratner et al. 2017a) in conjunction with active learning (Gonsior et al. 2020, Nashaat and Miller 2021) to address these challenges (see first two panels of Figure 2). We give an overview of our approach in what follows.

3.1. 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. The sensitive nature of the data (i.e., containing personal identifiable information such as names and contact information)

[3] A random manual labeling exercise of 1000 posts resulted in finding 0 recruitment posts, which provides the upper bound (i.e., with 95% probability, the mean number of recruitment posts is under 0.3%); the deterministic lower bound comes from the fact that our framework identified 6953 verified recruitment posts out of 13,568,130 total posts.

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 (Olsson 2009). However, as noted, we estimate around [99.7, 99.95]% (95% confidence interval) 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 thousands of randomly chosen posts to find a handful of recruitment instances with reasonable likelihood. More advanced sampling approaches – e.g., sampling from clusters of the data (Dligach and Palmer 2011) or dense regions (McCallumzy and Nigamy 1998), or maximizing diversity in the sample (Yang et al. 2015) – 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), which have a high tendency to overfit training data and are therefore data hungry (Chen and Lin 2014).

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 (Mikolov et al. 2013). 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 (Rehurek and Sojka 2010, Gensim 2021), 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 (Rong 2014). Following standard pre-processing techniques in natural language processing (Symeonidis et al. 2018), we drop stop words (e.g., 'the', 'is', 'but') and lemmatize the vocabulary (e.g., 'caring' would be converted ‘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’

Figure 3 List of expert-identified recruitment terms used to inform weak learners.

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, both of which are standard choices in the literature (see, e.g., Pennington et al. 2014).

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 Figure 3.

As with traditional weak supervised learning (Ratner et al. 2017a), 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 (Snorkel 2021), 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 (the remaining are sales-related posts). Note that this corresponds to 22% of the labels being positive, compared to only [0.05, 0.3]% (95% confidence interval) of the labels being positive on a random subsample of our dataset. However, this training dataset is biased based on the knowledge of domain experts and only uncovered 3 types of recruitment categories. Therefore, we proceed to the active learning stage to unmask the broader variation in recruitment categories across geographies.

3.2. Active Learning

The process designed for generating a training set provides us with initial well-balanced training data. 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 (Settles 2012, 2009). 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) 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 (Zhu et al. 2010).

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 categories (see full list in Table 3 in Appendix D), far outperforming the initial expert-identified vocabulary which only identified 3 types of recruitment categories.

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:

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

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 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. 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:

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

$\chi(\cdot)$ 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, $f(x) = 0$, or clearly recruitment, $f(x) = 1$) 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:

$$\Delta = \{x \in X_0 \mid 0.4 \leq f(x) \leq 0.8\} \tag{2}$$

$$\eta = \{x \in X_0 \mid f(x) > 0.8\}. \tag{3}$$

Δ captures uncertain posts, while η captures likely-recruitment posts. The upper and lower cutoffs (0.8 and 0.4, respectively) are tuning parameters that were chosen to optimize the recall of likely-recruitment posts under a given labeling budget; see Appendix B.2 for details.

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

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

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

$$N(\ell) = \frac{|\,\Delta \cap V(\ell)\,|^{\gamma}}{|\,\eta \cap V(\ell)\,| + 1}. \tag{4}$$

$N(\ell)$ captures the extent to which we distinguish potential recruitment categories 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.[4]

Higher values of γ will upweight locations with a larger volume of uncertain posts. We evaluate $\gamma \in \{1, 2\}$ in simulation (see detailed discussion and comparison in §4); we find $\gamma = 2$ works best for our use case with Tellfinder, which is what we use for our results on the deep web dataset.

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

$$M(e) = \frac{|\,\Delta \cap T(e)\,|^{\gamma}}{|\,\eta \cap T(e)\,| + 1}. \tag{5}$$

$M(e)$ 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.

Similar to node uncertainty, higher values of γ will upweight edges with a larger volume of uncertain 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:

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

We test $\gamma \in \{1, 2\}$ on synthetic data (see §4), and find that $\gamma = 2$ performs best for our proposed network discovery objective. Thus, our final results on deep web data (§5) use $\gamma = 2$.

[4] We add a small constant, 1, to the denominator to ensure that it is nonzero.

Active Learning Strategy. 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 $x \in X_0$ (several predictive models were compared prior to selecting the deep neural net architecture used, please see Appendix B.1 for details). 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 $x \in X_0 \cap \eta^c$ using a weighted combination of these prioritization metrics and choose the top $\approx 4,000$ 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.[5] Finally, we then re-compute our set of unlabeled likely-recruitment posts η; we stop the active learning process when this set is empty (see Algorithm 1).

Algorithm 1 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 $\eta \neq \varnothing$ **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 $x \in X_0 \cap \eta^c$
    Compute 'post rank' $R(x)$ using a weighted combination of $N(x), M(x)$ and $\chi(x)$ for every remaining $x \in X_0 \cap \eta^c$
    Sort remaining posts $x \in X_0 \cap \eta^c$ by descending order of $R(x)$
    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 set of unlabeled likely recruitment posts η
end

[5] The process primarily focuses on labeling posts as recruitment or sales. However, when finding a recruitment post, we additionally assign a recruitment category label. If a recruiting post didn't fit any of the already-identified categories, a new category was dynamically added with input from domain experts.

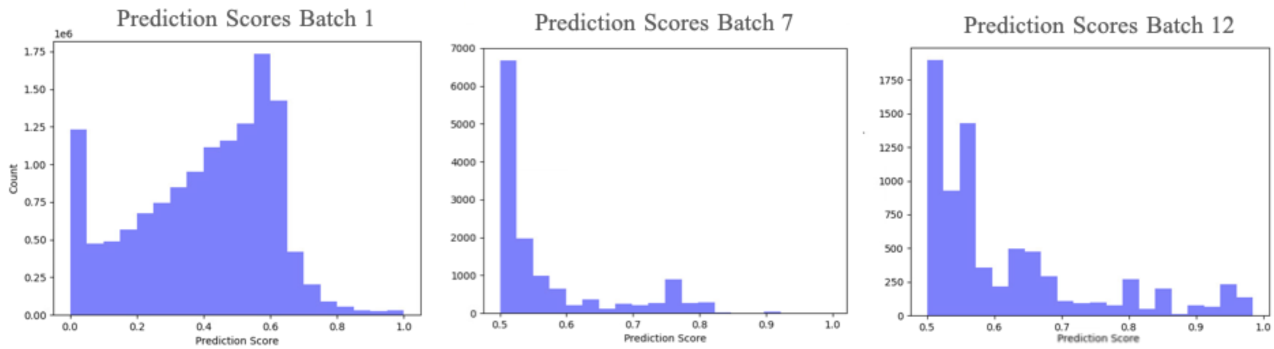


Figure 4 **Prediction scores from our model on the unlabeled data for three batches in the active learning process. Note that the y-axis and x-axis are different across plots — for Batch 7 and 12, we only show prediction scores above 0.5 to ensure readability (since very few posts have even a reasonable likelihood of being recruitment-related as the predictive model begins to converge).**

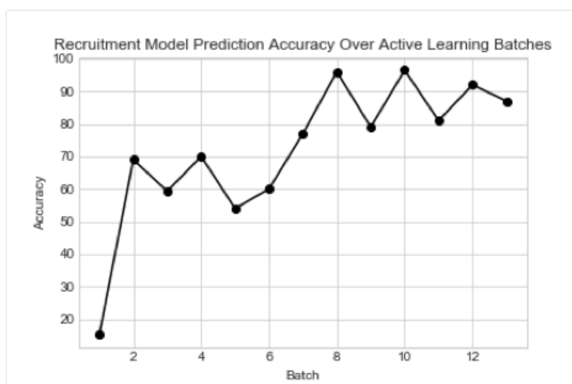


Figure 5 **Accuracy of the predictive model f across 13 batches of the active learning process.**

We ran 13 batches of active learning. Figure 4 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 4). Figure 5 shows how the accuracy of f evolves with each additional batch; note that its performance asymptotes, suggesting convergence.

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 categories, our active learning process uncovered 27 recruitment categories (see Appendix D), 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 categories we identify in smaller cities (where we actually find recruitment dominates).

Network Creation. 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 2). Along with the locations of the posts, this allows us to identify the geographic network connections underlying 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.

4. Synthetic Experiments

Before diving into the deep web data, we compare the performance of our active learning strategy against traditional active learning methodologies. Typically, active learning algorithms are evaluated based on model metrics (accuracy, recall, and precision). However, as discussed earlier, we have two additional goals:

1. Efficiently uncover recruitment-to-sales network pathways in our deep web data, and
2. Identify varying types of deceptive recruitment activity used across locations.

Therefore, in addition to typical model performance metrics, we introduce three new metrics for evaluating and comparing each algorithm. We compare our approach against both a basic active learning process and an active learning algorithm (Anahideh et al. 2022) designed for fair learning. With a limited budget, our algorithm performs best at discovering the underlying network structure and uncovering varying types of recruitment. The remainder of this section details our three new metrics, the key approaches we compared, and the results on synthetic data.

4.1. Synthetic Dataset Construction

We create a synthetic dataset that aims to mimic our Tellfinder dataset, in order to simulate and compare different active learning procedures. Each post generated in our synthetic dataset will have three components: a node assignment (“origin” location), an edge assignment (e.g., connection to another location representing a recruitment-to-sales pathway) and post content. Nodes and edges are assigned probabilistically based on the volumes of activity on nodes and edges found in our TellFinder dataset within the United States and Canada in order to maintain the same underlying network structure as our original data. We randomly generate 100-dimensional real-valued features representing posts,[6] binary labels (recruitment vs. sales), as well as clusters (recruitment categories); details on how this data is generated are provided in Appendix E.1. Thus, we construct a synthetic dataset with approximately 450,000 posts (that have known labels and network associations) associated with 697 nodes. We split the dataset into a training dataset (≈300,000 posts) for conducting the active learning simulations and a test dataset (≈150,000 posts) for evaluating/comparing the models created during the active learning procedures. This allows us to effectively evaluate the network discovery performance of different active learning approaches against the known ground truth network.

4.2. Methods Compared

We compare the following active learning schemes:

1. *Normalized Network Uncertainty Metrics (NUM):* Our proposed approach, which combines prediction uncertainty with metrics on node and edge uncertainty to improve network discovery. NUM prioritizes posts that are located on nodes or edges with a high share of uncertainty in predictions, thereby focusing on nodes where there is uncertainty regardless of the volume of posts at that node. Posts are prioritized based a weighted combination of prediction, node, and edge uncertainty, with weights 0.75, 0.125, and 0.125, respectively, chosen to balance model performance with network discovery using 10-fold cross-validation. Prediction, node, and edge uncertainties are computed according to (1), (4), and (5), respectively, with $\gamma = 1$.

2. *Scaled Network Uncertainty Metrics (SNUM):* The same as NUM, except using $\gamma = 2$ instead of $\gamma = 1$. We also use the same weights to balance prediction, node, and edge uncertainty.[7]

3. *Base Active Learning (BaseAL):* Base active learning leverages the level of uncertainty for the model prediction (e.g., how close the prediction is to being a random draw). Thus, posts are prioritized for labeling purely based on prediction uncertainty, not including any network discovery objectives. Uncertainty is computed using $\chi(x)$ as defined in (1).

4. *Regional Fairness and Misclassification (FAL):* Finally, we consider an approach by Anahideh et al. (2022), a recently proposed method that incorporates fairness into the active learning process. This algorithm prioritizes posts based on a weighted average of the Shannon entropy (to reduce misclassification error) and a post's potential improvement in expected fairness across groups. Details of the algorithms are provided in Appendix E.2.

We take the fairness metric to be the well-known notion of demographic parity (Hardt et al. 2016) — in our setting, this implies that we should identify recruitment posts at similar rates

[6] We directly generate real-valued feature vectors instead of unstructured text (as in our deep web data), allowing us to study the performance of active learning without the added complexity of converting raw text into real-valued feature vectors via word embeddings.

[7] Note that both NUM and SNUM balance model performance with network discovery in the same way.

across regions. This metric ensures that we uncover recruitment activity in less populated areas (an important ingredient of our results, as we will discuss in §5, since less-populated areas tend to be recruitment hotspots). We define our regions as the following: Canada, USA West, USA South, USA Northeast, and USA Midwest.[8]

Remark 1. Other advances in active learning include improving sampling diversity (Citovsky et al. 2021, Ash et al. 2019) and incorporating network features (Bilgic et al. 2010). However, these approaches are not applicable to our setting. Methods that leverage sampling diversity metrics rely on clustering the data to query labels across clusters that are closest to the cluster mean; this would lead to noisy and inefficient labeling with highly imbalanced data (i.e., most clusters will capture different types of sales posts and will not be relevant for finding recruitment activity). Methods that incorporate network features assume a known network structure.

4.3. Experiment Design

We run each active learning procedure with the same starting batch for 20 batches. For each batch, we run 10-fold cross-validation using the training data available for that batch and compute label prediction scores for each post. We compute the specific active learning prioritization metrics across the unlabeled data and then select the top 100 posts to label,[9] and add those posts to the training dataset. After 20 batches, we stop the active learning process. We repeat this process for each technique for 50 runs and compute confidence intervals around key performance metrics.

4.4. Network Recovery Metrics

We define the following three new metrics, motivated by counter-trafficking needs:

1. *Edge investment return:* As discussed earlier, connecting recruitment activity to sex sales conducted by the same entity allows us to infer likely recruitment-to-sales trafficking routes of criminal entities—this can inform partnerships between law enforcement agencies at either end of the trafficking supply chain, significantly improving the efficiency of counter-trafficking efforts (Heilemann and Santhiveeran 2011, Hodge and Lietz 2007, Johnson 2012, Jones et al. 2007, Roby 2005). However, such partnerships are costly to create and maintain. Thus, we aim to compare how well each method uncovers relevant edges in the trafficking network. We define "edge investment return" as the aggregate amount of recruitment-to-sales activity that could be uncovered if law enforcement were to invest resources in partnering and coordinating across the top 50 edges discovered by a

[8] We found that more granular regions (e.g., city-level) resulted in worse performance and prohibitively long runtimes (due to the numerous fairness computations for evaluating the potential fairness improvement of each post).

[9] Note that we label fewer posts per batch since our synthetic data has simple numeric features with a relatively low signal-to-noise ratio, compared to our Tellfinder text dataset that has a high signal-to-noise ratio (as evidenced by the number of samples needed for the model performance to converge).

particular method. In more detail, given the set E_x of edges associated with a particular post x, we define the set of all posts associated with a pair of locations on a given edge e as

$$P(e) = \{x \in X \mid e \in E_x\}.$$

Then, we define the inferred set of recruitment posts identified by a particular method m as:

$$\eta_m = \{x \in X \mid f_m(x) > \alpha\}. \tag{6}$$

where α represents a prediction cutoff for considering a post to be a positive label.[10] Thus,

$$A_m(e) = |\eta_m \cap P(e)|,$$

measures the inferred recruitment-to-sales activity found on a particular edge e.[11] We thus rank edges for a method m by computing the inferred recruitment-to-sales activity level $A_m(e)$ for each edge e, and then sorting the edges in descending order of $A_m(e)$ — i.e., letting $e_1^m, \ldots, e_k^m$ be the edges sorted in descending order of $A_m(e)$, we have $A_m(e_1^m) \geq A_m(e_2^m) \geq \ldots \geq A_m(e_k^m)$. We now sum the *true* recruitment-to-sales activity (as measured by our known ground truth network) for the top 50 ranked edges. Thus, the edge investment return is $\bar{A}_m = \sum_{i=1}^{50} A^*(e_i^m)$, where

$$A^*(e) = |\eta^* \cap P(e)|$$
$$\eta^* = \{x \in X \mid f^*(x) = 1\},$$

and where $f^*(x)$ is the ground truth value of whether x is a recruitment post (so $A^*(e)$ is the true recruitment-to-sales activity of edge e).

2. *Top edge overlap:* While edge investment return focused on the total volume of activity found, we next study the accuracy of each method in identifying the riskiest edges in the true underlying network. To this end, we compare the top 50 edges identified by each method against the actual top 50 edges in the true network. In particular, letting $e_1^m, e_2^m, \ldots, e_k^m$ be the inferred high-risk edges (sorted in descending order of $A_m(e)$, i.e., $A_m(e_1^m) \geq \ldots \geq A_m(e_k^m)$), and letting $e_1^*, e_2^*, \ldots, e_k^*$ be the true high-risk edges (sorted in descending order of $A^*(e)$), we report the number of true high-risk edges identified by a method m as $|\{e_1^m, \ldots, e_{50}^m\} \cap \{e_1^*, \ldots, e_{50}^*\}|$.

3. *Cluster recovery fairness:* As we will show in §5, there is significant variation in the recruitment types/categories used to target vulnerable populations in different locations. Thus, to ensure fairness, it is important that our methods are able to recover recruitment activity across different

[10] For each method, α is chosen to optimize the edge investment return using 10-fold cross-validation.

[11] This measure also captures recruitment-recruitment activity, but this kind of activity is minimal (since recruitment is rare) and also of interest to law enforcement agencies based on our conversations.

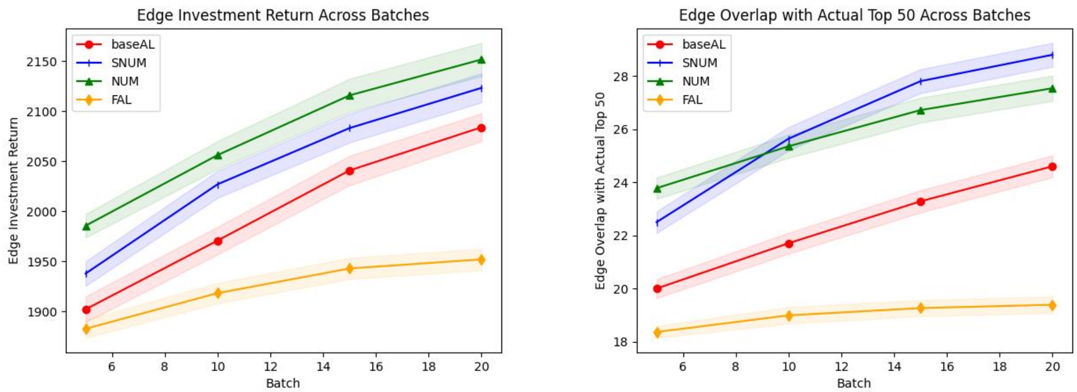


Figure 6 **Edge investment return (left) and top edge overlap (right) for each method. NUM and SNUM outperform BaseAL and FAL across all labeling budgets (number of batches).**

recruitment types (or clusters in our synthetic data). To this end, we use the classic Gini coefficient to capture inequality in recall across recruitment clusters. In particular, letting τ denote a recruitment type and $\nu_\tau = \{x \in X \mid x \text{ has type } \tau\}$, we compute $z_{m,\tau}$ as method m's recall rate among posts in ν_τ (measured with respect to the ground truth recruitment labels). We report the Gini coefficient of $\{z_{m,\tau}\}_\tau$; lower values represent fair recovery across recruitment types.

4.5. Results on Synthetic Data

Figure 6 shows the edge investment return and the fraction of top edges recovered for each method. We see that SNUM and NUM outperform the other methods across all labeling budgets (i.e., number of batches). Our results demonstrate that the inclusion of network discovery metrics in the active learning process enables us to better identify high-risk recruitment-to-sales pathways to support counter-trafficking efforts. In contrast, BaseAL and FAL sample with the primary goal of improving model accuracy (and additionally regional fairness for FAL), rather than pursuing potentially high-risk edges.

Next, we identify 14 distinct recruitment "clusters" in the synthetic dataset, each meant to represent different recruitment types. Figure 7 shows both a two dimensional visualization of the clusters,[12] as well as the cluster recovery fairness metrics for each method (recall that a lower Gini coefficient implies fairer recovery). Again, we see that SNUM and NUM have significantly fairer recovery across all prediction cutoff thresholds used for the model. In other words, SNUM and NUM are not only better at uncovering edges with high recruitment-to-sales activity, but also obtain fair coverage across different types of recruitment tactics. Interestingly, this happens even though SNUM and NUM do not explicitly prioritize discovering different recruiting types.

[12] We visualize the clusters by projecting post features onto two dimensions using the uniform manifold approximation and projection (UMAP) dimensionality reduction technique (McInnes et al. 2018).

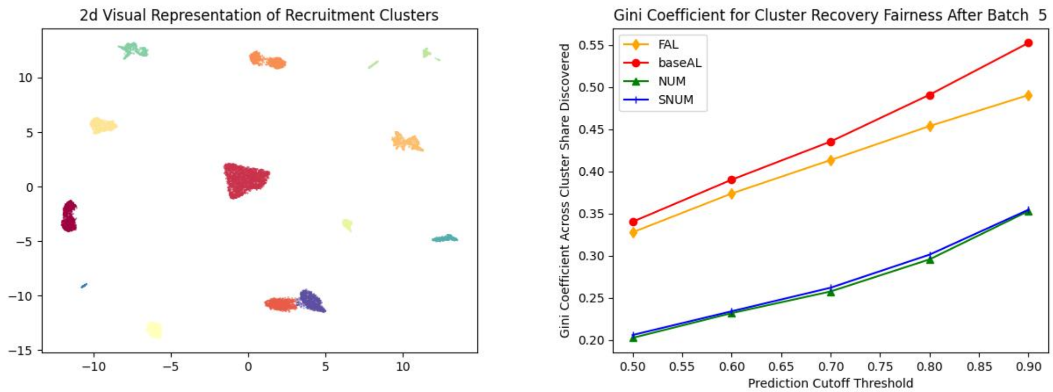


Figure 7 **Two-dimensional visualization of recruitment post clusters in our synthetic dataset (left). Cluster recovery fairness (right), as measured by the Gini coefficient, for each method. NUM and SNUM exhibit lower inequality than BaseAL and FAL across all prediction cutoff thresholds.**

Lastly, we investigate whether there is a tradeoff between network discovery (the focus of our metrics) and improving the precision/recall of the model f (the focus of classic active learning). To this end, we look at the Area Under the Receiver Operator Characteristic Curve (AUC) — a popular metric for assessing overall model precision/recall — for each method (see Figure 8). When evaluating model performance across the entire network, we see that NUM and BaseAL perform the best across all labeling budgets. However, in our application, the end user is primarily interested in maximizing model performance among *high-risk* portions of the network (here, we measure high-risk as the top 50 recruitment-to-sales edges, formally defined earlier). This is where law enforcement seeks to direct costly partnerships and interventions; arguably, due to the stringent resource constraints faced by law enforcement agencies, effort is restricted to high-risk regions, so the predictive model is only informative of decisions in those regions. To this end, we evaluate AUC for posts associated with the top 50 edges in the ground truth network (Figure 8). We see that NUM performs the best for very small labeling budgets, but SNUM performs significantly better for moderate labeling budgets. Thus, depending on the labeling budget, NUM or SNUM should be used when the goal is to accurately label posts in high-risk regions of the network.

4.6. Discussion

In summary, by leveraging network discovery metrics in the active learning process, both NUM and SNUM significantly improve our ability to identify high-risk edges of the trafficking network. In addition, NUM and SNUM recover these edges much more fairly across recruitment types. NUM achieves similar model performance (measured by AUC) across the entire network relative to base active learning, while SNUM maximizes model performance among high-risk portions of

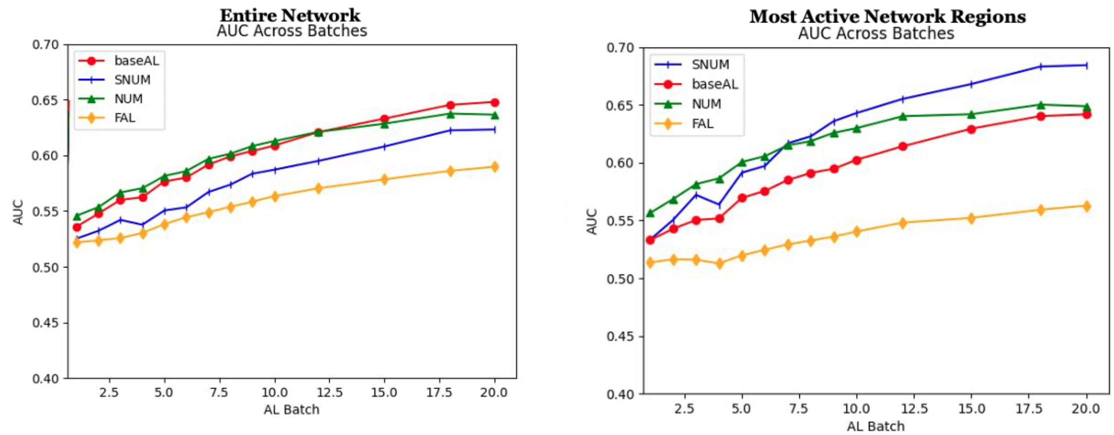


Figure 8 **AUC for each method across the entire network (left) and restricted to posts in the top 50 high-risk edges of the network (right). NUM and BaseAL perform best across the entire network, while SNUM performs best with moderate labeling budgets in the high-risk regions of the network.**

the network. Thus, we find strong evidence for using NUM or SNUM when the objective is not simply model performance, but also network discovery.

Recall that NUM and SNUM are identical by design, except SNUM further upweights nodes/edges with larger volumes of uncertain posts. This leads SNUM to perform more exploration in early batches, leading to improved performance in high-risk regions of the network in later batches. Thus, SNUM should be used when a user has at least a moderate labeling budget and is primarily interested in accurately uncovering activity in high-risk regions; in contrast, NUM should be used when a user has a very small labeling budget, or is interested in improving AUC across the entire network (regardless of risk/activity). Notably, both algorithms achieve comparably fair coverage across recruitment types.

For our application with the TellFinder Alliance, we find that SNUM is more appropriate since we have a moderate labeling budget, and our partners were primarily interested in predictions informing high-risk recruitment-to-sales pathways (while preserving fairness across recruitment types) to help target interventions. Therefore, our results on deep web data in the remainder of the paper rely on using SNUM.

5. Results on Deep Web Data

We now apply our machine learning framework to our deep web dataset to extract a network view of trafficking risk in online commercial sex supply chains.

5.1. Recruitment Activity

Figure 9 shows the global map of recruitment hotspots and the types of recruitment categories identified in our data. Note that recruitment posts in the ‘escort’ category could indicate potential involvement with selling sexual services, while posts in all other categories have no such indication. As shown in Figure 9a, 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 (see discussion in §6).

We observe significant geographic variation in the approaches used to recruit victims (see Figures 9a and 9b); the full list of recruitment types is in Appendix D. For example, within the United States, individuals are primarily targeted through porn, modeling, and adult entertainment in smaller cities; in larger cities, they are primarily targeted through personal ads, sugar parent requests, and escort services (see Appendix F). More globally, victims are targeted 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 (Murphy 2016). These results can be used to tailor such educational programs towards the currently popular recruitment approaches in those specific locales.

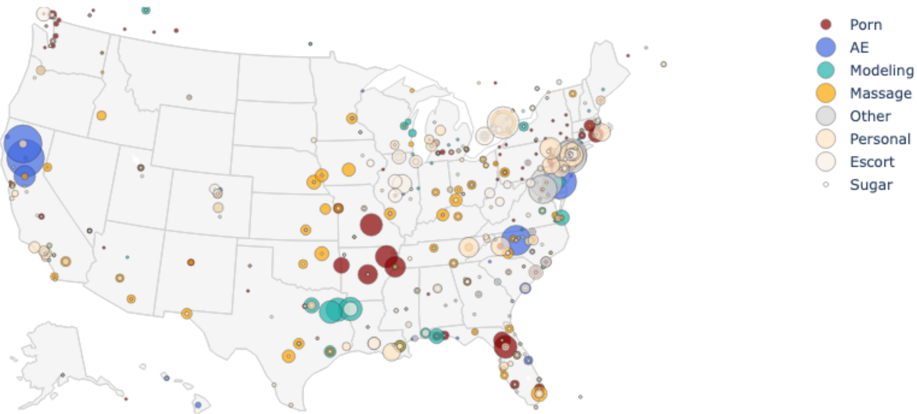
We also construct recruitment-recruitment networks, creating an undirected edge between a pair of locations if they each have a high volume of recruiting posts (at least 150) by the same entity. Figure 9c shows the resulting network within the United States. We observe that many recruiters operate in multiple locations spanning large distances, suggesting a highly organized effort.

5.2. Inferring Human Trafficking Risk

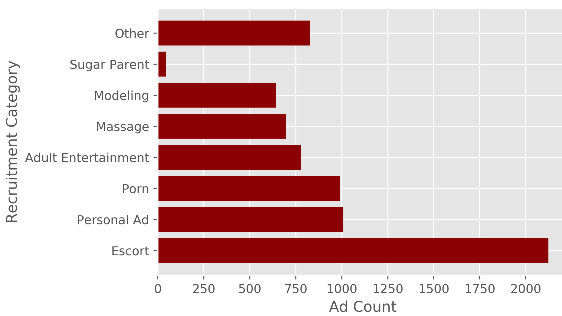
As discussed earlier, linking likely deceptive (non-sex) recruitment offers to commercial sex sales by the same entity strongly suggests that trafficking may have occurred. Figure 10a shows an example connection between an identified recruitment post and two sales posts with shared meta data; although 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.

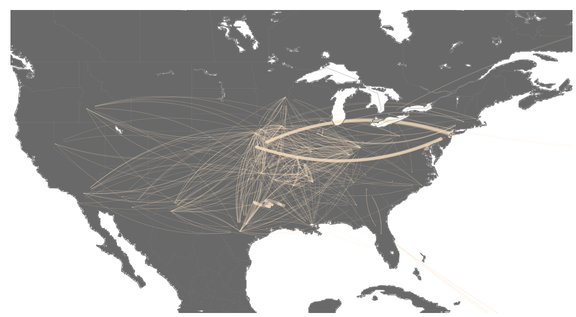
To study trafficking risk at scale, we construct recruitment-to-sales pathways: we create a directed edge between a pair of locations if there is a recruiting post and a sales post with shared meta data. Figure 10b shows the resulting commercial sex supply chain, restricting to edges that have at least 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,



(a) Recruitment Activity by Category Uncovered by Active Learning Algorithm



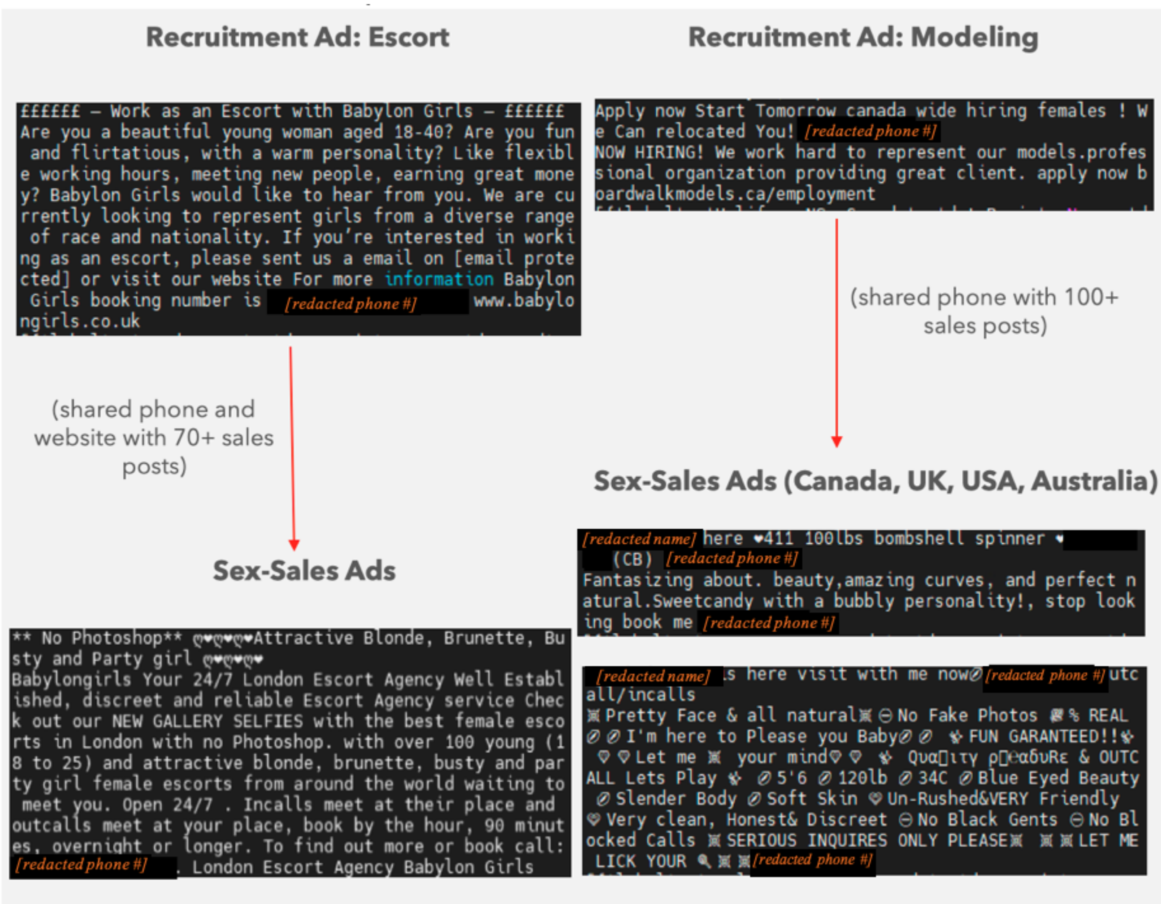


(b) Types of Recruitment (c) Connections amongst Recruiters

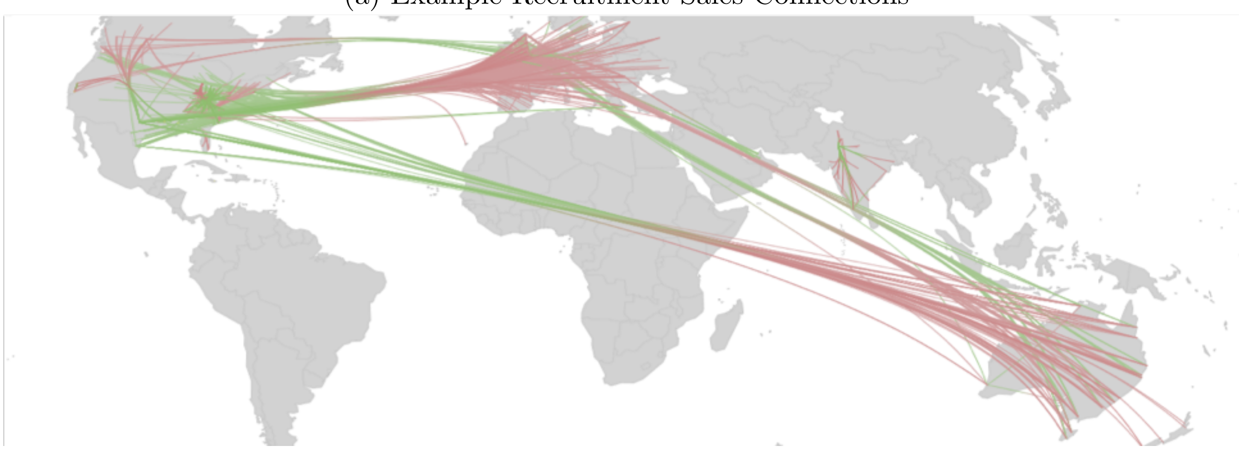
Figure 9 **The top panel (a) shows recruitment hotspots and categories identified in our data. Larger markers indicate more posts. The bottom left panel (b) shows the histogram of recruitment posts by category across the world. The bottom right panel (c) shows the recruitment-recruitment network in the 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).**

which targets network discovery in addition to the traditional objective of improving classification accuracy.

Figure 11 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 11a) and India (Figure 11b).

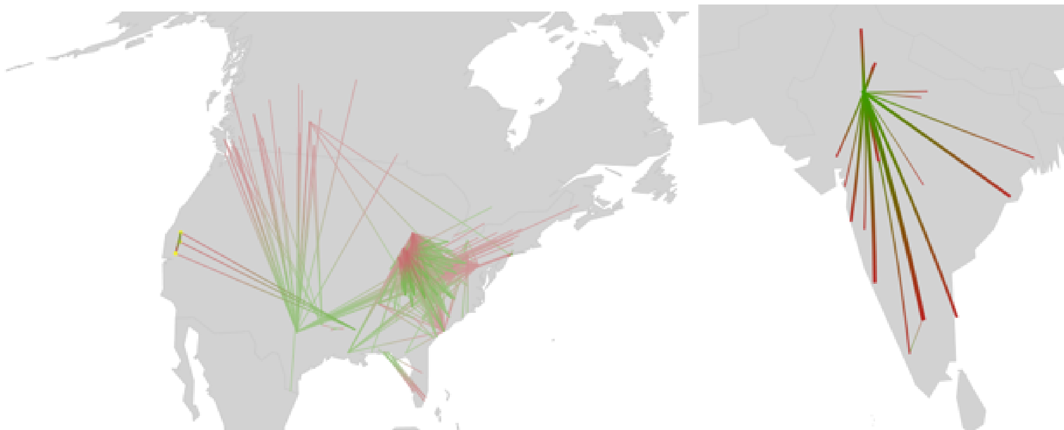


(a) Example Recruitment-Sales Connections



(b) Recruitment to Sex-Sales Pathways Unmasked

Figure 10 **The top panel (a) shows example escort and modeling recruitment posts uncovered 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) unmasked from our algorithm. Network is restricted to edges with at least 100 occurrences.**



(a) USA Recruitment- Sex Sales Network

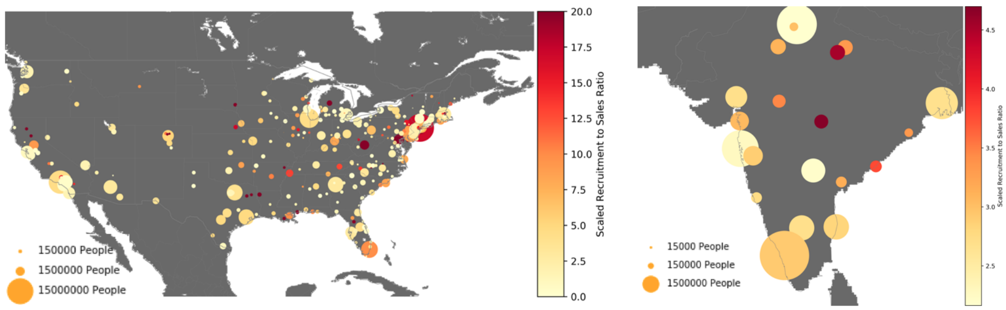
(b) India Recruitment- Sex Sales Network

Figure 11 **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.**

5.3. 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 12a 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 identified 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 Appendix F. 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



(a) USA Recruitment- Sex Sales Pressure Ratios (b) India Recruitment- Sex Sales Pressure Ratios

Figure 12 **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).**

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.

5.4. 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. Details are provided in Appendix F. 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.

6. Discussion

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 likely 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" (FBI 2021). 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 (FBI 2019), 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 (Shively et al. 2012). 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 that can 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 (Orme and Ross-Sheriff 2015), as well as rehabilitating victims after their trafficking experience (Rafferty 2008). 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 (Jorgensen and Sandoval 2019), and 93% of victims in Canada are Canadian citizens (Lopez-Martinez 2020). 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 categories 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; these methods have already shown success identifying sex sales ads by the same organization (Li et al. 2018, Dubrawski et al.

2015). We note that it is unlikely that criminals will respond with these shifts in the near term. Finally, while deep web data provides a significant opportunity to scale the collection of information, it may fail to provide adequate coverage of some vulnerable populations. For instance, half of the cases reviewed by the UNODC in 2020 used the internet (Kangaspunta et al. 2020), but there is still a significant amount of trafficking activity conducted offline (e.g., through word-of-mouth). Relatedly, our choice of websites (informed by law enforcement partners) as well as limitation to the English language may limit visibility of illicit activity occurring elsewhere or in local languages. As discussed in Section 2, a promising direction of future work is adapting our approach to other languages to improve global coverage. Thus, it is important to note that any insights from our approach should complement rather than replace traditional leads (e.g., survivor interviews, prior case outcomes, etc.), which may provide better coverage over vulnerable populations that are underrepresented by our analysis.

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.

References

Albright E, D'Adamo K (2017) Decreasing human trafficking through sex work decriminalization. *AMA journal of ethics* 19(1):122–126.

Anahideh H, Asudeh A, Thirumuruganathan S (2022) Fair active learning. *Expert Systems with Applications* 199:116981.

Androff DK (2011) The problem of contemporary slavery: An international human rights challenge for social work. *International Social Work* 54(2):209–222.

Ash JT, Zhang C, Krishnamurthy A, Langford J, Agarwal A (2019) Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671* .

Baird K, Connolly J (2023) Recruitment and entrapment pathways of minors into sex trafficking in canada and the united states: A systematic review. *Trauma, Violence, & Abuse* 24(1):189–202.

BEA (2018) U.s. bureau of economic analysis. URL `https://apps.bea.gov/regional/downloadzip.cfm`.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Bilgic M, Mihalkova L, Getoor L (2010) Active learning for networked data. *Proceedings of the 27th international conference on machine learning (ICML-10)*, 79–86.

BLS (2018) U.s. bureau of labor statistics. URL `https://www.bls.gov/lau/#tables`.

Bouche V, Crotty SM (2018) Estimating demand for illicit massage businesses in houston, texas. *Journal of human trafficking* 4(4):279–297.

Census (2018) U.s. census. URL `https://data.census.gov/cedsci/`.

Chan H, Tran-Thanh L, Wilder B, Rice E, Vayanos P, Tambe M (2018) Utilizing housing resources for homeless youth through the lens of multiple multi-dimensional knapsacks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 41–47.

Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. *IEEE access* 2:514–525.

Citovsky G, DeSalvo G, Gentile C, Karydas L, Rajagopalan A, Rostamizadeh A, Kumar S (2021) Batch active learning at scale. *Advances in Neural Information Processing Systems* 34:11933–11944.

Dank ML, Khan B, Downey PM, Kotonias C, Mayer D, Owens C, Pacifici L, Yu L (2014) Estimating the size and structure of the underground commercial sex economy in eight major us cities .

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Diaz M, Panangadan A (2020) Natural language-based integration of online review datasets for identification of sex trafficking businesses. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, 259–264 (IEEE).

Dligach D, Palmer M (2011) 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*, 6–10.

Do CB, Ng AY (2005) Transfer learning for text classification. *Advances in Neural Information Processing Systems* 18:299–306.

Dubrawski A, Miller K, Barnes M, Boecking B, Kennedy E (2015) Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking* 1(1):65–85.

FBI (2016) U.s. department of justice. URL `https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-8/table-8.xls/view`.

FBI (2019) Operation independence day,. URL `https://www.fbi.gov/news/stories/operation-independence-day-2019`.

FBI (2021) Human trafficking,. URL `https://www.fbi.gov/investigate/violent-crime/human-trafficking`.

Flynn C, Alston M, Mason R (2014) Trafficking in women for sexual exploitation: Building australian knowledge. *International Social Work* 57(1):27–38.

Gensim (2021) Gensim 4.0.1. URL `https://www.snorkel.org/`.

Gezinski LB, Gonzalez-Pons KM (2022) Sex trafficking and technology: A systematic review of recruitment and exploitation. *Journal of Human Trafficking* 1–15.

Gonsior J, Thiele M, Lehner W (2020) Weakal: Combining active learning and weak supervision. *International Conference on Discovery Science*, 34–49 (Springer).

Hall E, Dickson C, Schroh D, Wright W (2015) Tellfinder: Discovering related content in big data (VIS).

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29.

Heilemann T, Santhiveeran J (2011) 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* 20(1):57–76.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural computation* 9(8):1735–1780.

Hodge DR, Lietz CA (2007) The international sexual trafficking of women and children: A review of the literature. *Affilia* 22(2):163–174.

HUD (2019) Us department of housing and urban development. URL `https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/`.

ILO (2017) Human trafficking by the numbers.

Johnson BC (2012) Aftercare for survivors of human trafficking. *Social Work & Christianity* 39(4).

Jones L, Engstrom DW, Hilliard T, Diaz M (2007) Globalization and human trafficking. *J. Soc. & Soc. Welfare* 34:107.

Jorgensen S, Sandoval P (2019) Experts: Trump's tape bound women trafficking claim is misleading,. *CNN* URL `https://www.cnn.com/2019/01/27/us/human-trafficking-fact-check/index.html`.

Kangaspunta K, Sarrica F, Serio G, Kelly W, Samson J, Wills C (2020) Global report on trafficking in persons 2020 .

Kaya YB, Maass KL, Dimas GL, Konrad R, Trapp AC, Dank M (2022) Improving access to housing and supportive services for runaway and homeless youth: Reducing vulnerability to human trafficking in new york city. *arXiv preprint arXiv:2202.00138* .

Kejriwal M, Kapoor R (2019) Network-theoretic information extraction quality assessment in the human trafficking domain. *Applied Network Science* 4(1):1–26.

Keskin BB, Bott GJ, Freeman NK (2021) Cracking sex trafficking: Data analysis, pattern recognition, and path prediction. *Production and Operations Management* 30(4):1110–1135.

Kessler G (2018) Has the sex-trafficking law eliminated 90 percent of the sex trade. *Washington Post* .

Konrad R, Maass KL, Trapp AC (2020) A perspective on how to conduct responsible anti-human trafficking research in operations and analytics. *arXiv preprint arXiv:2006.16445* .

Konrad RA, Trapp AC, Palmbach TM, Blom JS (2017) Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications. *European Journal of Operational Research* 259(2):733–745.

Kosmas D, Melander C, Singerhouse E, Sharkey TC, Maass KL, Barrick K, Martin L (2022) Generating synthetic but realistic human trafficking networks for modeling disruptions through transdisciplinary and community-based action research. *arXiv preprint arXiv:2203.01893* .

Kotrla K (2010) Domestic minor sex trafficking in the united states. *Social work* 55(2):181–187.

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5(4):221–232.

Laczko F (2002) Human trafficking: the need for better data. *Migration Information Source* 1:61–80.

Latonero M (2011) Human trafficking online: The role of social networking sites and online classifieds. *Available at SSRN 2045851* .

Learn S (2021) Make classification. URL `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html`.

Li L, Simek O, Lai A, Daggett M, Dagli CK, Jones C (2018) Detection and characterization of human trafficking networks using unsupervised scalable text template matching. *2018 IEEE International Conference on Big Data (Big Data)*, 3111–3120 (IEEE).

Li R, Tobey M, Mayorga M, Caltagirone S, zaltın O (2021) Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Available at SSRN 3982796* .

Lin M, Chen Q, Yan S (2013) Network in network. *arXiv preprint arXiv:1312.4400* .

Liu G, Guo J (2019) Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337:325–338.

Lopez-Martinez M (2020) Sex trafficking still a prevalent issue across canada, advocates and police say. *CTVNews* URL `https://www.cnn.com/2019/01/27/us/human-trafficking-fact-check/index.html`.

Maass KL, Trapp AC, Konrad R (2020) Optimizing placement of residential shelters for human trafficking survivors. *Socio-Economic Planning Sciences* 70:100730.

Martin L, Pierce A, Peyton S, Gabilondo AI, Tulpule G (2014) Mapping the market for sex with trafficked minor girls in minneapolis: Structures, functions, and patterns .

McCallumzy AK, Nigamy K (1998) Employing em and pool-based active learning for text classification. *Proc. International Conference on Machine Learning (ICML)*, 359–367 (Citeseer).

McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* .

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Murphy LT (2016) Labor and sex trafficking among homeless youth. *A Ten City Study (Executive Summary)* .

Nashaat M, Miller J (2021) Improving news popularity estimation via weak supervision and meta-active learning. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2679.

NCES (2017) National center for education statistics. URL `https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7`.

Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378* .

Office GA (2006) Human trafficking: Better data, strategy, and reporting needed to enhance us antitrafficking efforts abroad. *Trends in Organized Crime* 10:16–38.

Okech D, Choi YJ, Elkins J, Burns AC (2018) Seventeen years of human trafficking research in social work: A review of the literature. *Journal of evidence-informed social work* 15(2):103–122.

Olsson F (2009) A literature survey of active machine learning in the context of natural language processing .

Orme J, Ross-Sheriff F (2015) Sex trafficking: Policies, programs, and services. *Social work* 60(4):287–294.

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Potocky M (2010) The travesty of human trafficking: A decade of failed us policy. *Social Work* 55(4):373–375.

Proximity (2009) Proximity. URL `http://proximityone.com/metro_healthinsurance.htm`.

Rabbany R, Bayani D, Dubrawski A (2018) Active search of connections for case building and combating human trafficking. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2120–2129.

Raets S, Janssens J (2019) Trafficking and technology: Exploring the role of digital communication technologies in the belgian human trafficking business. *European Journal on Criminal Policy and Research* 1–24.

Rafferty Y (2008) The impact of trafficking on children: Psychological and social policy perspectives. *Child Development Perspectives* 2(1):13–18.

Rajapakse T (2020) Simpletransformers. URL `https://simpletransformers.ai/`.

Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C (2017a) Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, 269 (NIH Public Access).

Ratner AJ, Ehrenberg HR, Hussain Z, Dunnmon J, Ré C (2017b) Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems* 30:3239.

Rehurek R, Sojka P (2010) Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (Citeseer).

Roby JL (2005) Women and children in the global sex trade: Toward more effective policy. *International Social Work* 48(2):136–147.

Roby JL, Vincent M (2017) Federal and state responses to domestic minor sex trafficking: the evolution of policy. *Social work* 62(3):201–210.

Romano A (2018) A new law intended to curb sex trafficking threatens the future of the internet as we know it. *Vox* .

Rong X (2014) word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* .

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11):2673–2681.

Settles B (2009) Active learning literature survey .

Settles B (2012) Active learning (synthesis lectures on artificial intelligence and machine learning)(morgan and claypool publishers, san rafael, ca) .

Shelters W (2021) Women's shelters. URL https://www.womenshelters.org/#state_list.

Shively M, Kliorys K, Wheeler K, Hunt D (2012) A national overview of prostitution and sex trafficking demand reduction efforts .

Smirnov NV (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* 2(2):3–14.

Snorkel (2021) Snorkel. URL https://www.snorkel.org/.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1):1929–1958.

Symeonidis S, Effrosynidis D, Arampatzis A (2018) A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications* 110:298–310.

TellFinder (2021) Tellfinder alliance: a global counter-human trafficking partner network, empowered by data,. URL https://www.tellfinderalliance.com/.

TensorFlow (2021) Tensorflow. URL https://www.tensorflow.org/.

UNODC (2020) Unodc human trafficking case law database .

Witte M (2018) The anti-trafficking movement needs better data to solve the problem, stanford researchers say. *Stanford News Service* URL https://news.stanford.edu/press-releases/2018/09/05/get-good-data-human-trafficking/.

Yang Y, Ma Z, Nie F, Chang X, Hauptmann AG (2015) Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision* 113(2):113–127.

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* .

Zhang C, Ré C, Cafarella M, De Sa C, Ratner A, Shin J, Wang F, Wu S (2017) Deepdive: Declarative knowledge base construction. *Communications of the ACM* 60(5):93–102.

Zhou ZH (2018) A brief introduction to weakly supervised learning. *National science review* 5(1):44–53.

Zhu J, Li L, Jones C (2019) Identification and detection of human trafficking using language models. *2019 European Intelligence and Security Informatics Conference (EISIC)*, 24–31 (IEEE).

Zhu J, Wang H, Hovy E, Ma M (2010) Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing (TSLP)* 6(3):1–24.

Appendix A: FOSTA-SESTA Impact

On April 11, 2018, the United States federal government passed two bills: FOSTA (Allow States and Victims to Fight Online Sex Trafficking Act) and SESTA (Stop Enabling Sex Traffickers Act). The combination of laws is referred to as FOSTA-SESTA, which made it illegal to knowingly assist, facilitate, or support sex trafficking, ultimately making website publishers responsible for any sex work (consensual or trafficking) advertised on their site (Romano 2018). In addition, on April 6, 2018, the Department of Justice announced the seizure and removal of Backpage.com, one of the top sites used for adult services transactions. These two events had a substantial impact on the commercial sex industry and online trafficking activity. However, they did not actually reduce online commercial sex activity; rather, advertising for the sex trade shifted to new websites (see, e.g., Kessler 2018), which made it even more important to provide updated trafficking risk models to law enforcement partners.

These events both occurred during the timeframe of our data sample. Therefore, we worked with the TellFinder Alliance to select the top sites most frequently used *after* April 2018 to ensure that our results are representative of the post FOSTA-SESTA environment (since data on sites used primarily before FOSTA-SESTA would show an outdated view of activity). Indeed, 89% of the posts in our data occurred after April 2018 (i.e., when activity dispersed from Backpage.com to new websites).

In addition, we compare the trafficking network we identified before and after these events. We find that 96% of the nodes, 90% of overall recruitment cities, and 88% of the top 50 recruitment cities that we identified are the same before and after FOSTA-SESTA. 30 new recruitment cities were added in the post FOSTA-SESTA time period, reflecting increased activity. The main recruitment categories we identified stayed the same, but there was an increase in recruitment for escort services, massage, and "Other" ads. Ads in the "Other" category were those that did not fit into any of the high-frequency recruitment categories — examples include cleaning services, secretarial services, etc (full list given in Table 3 in Appendix D, beginning with 'Non-specified agency' and thereafter). This shift indicates a larger diversity of deceptive online recruitment efforts on these websites post FOSTA-SESTA, highlighting the need for updated predictive models.

Appendix B: Model Selection

B.1. Predictive Model

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, but within this class, there are still 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) (Ratner et al. 2017b). 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 2). The first four models are built using Keras in Tensorflow (TensorFlow 2021). 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 (Lin et al. 2013). The final layer is a densely connected layer with a sigmoid activation function, which is useful for predicting probabilities (Nwankpa et al. 2018). 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 (Srivastava et al. 2014). We also include a bias initializer to help address the remaining data imbalance in our training set (Krawczyk 2016). 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 (Liu and Guo 2019). We test both a simple LSTM ("Model 3") (Hochreiter and Schmidhuber 1997), and a bi-directional LSTM ("Model 4") that leverages both the input sequence and a reversed copy in order to learn the whole context (Schuster and Paliwal 1997). Finally, we test transfer learning from two state-of-the-art language models, BERT (Devlin et al. 2018) and XLNET (Yang et al. 2019), using the Simple Transformers package (Rajapakse 2020). Transfer learning allows us to take advantage of pre-training on larger datasets and fine-tune a model to our particular classification task (Do and Ng 2005). The results of the 6 models tested are shown in Table 2.

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.

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%

Table 2 Different DNN architectures tested prior to active learning process.

B.2. Parameter Selection

Recall from Algorithm 1 that the upper and lower cutoffs used to define Δ and η in (2) and (3) affect the uncertainty score via the node and label uncertainties $N(x)$ and $M(x)$, respectively. Intuitively, as the upper cutoff increases, then Δ becomes larger and η becomes smaller; thus, $N(x)$ becomes relatively larger for posts associated with locations with many uncertain posts (i.e., $\Delta \cap V(\ell)$ is large). As a consequence, increasing this cutoff increases exploration (i.e., the ability to discover additional recruitment templates) since more uncertain posts are prioritized for labeling within our fixed labeling budget. On the other hand, increasing this cutoff decreases exploitation, since as η becomes smaller, we label fewer high-confidence posts. Thus, these cutoffs must be manually tuned to attain good performance (in this case, node-level recall on a held-out test set). In our simulations (which attempted to mimic the data-generating process for our dataset), this tradeoff is optimized when the cutoff is 0.8, so we use this value in our experiments (see Figure 13). A similar consideration leads to choosing the lower cutoff to be 0.4.

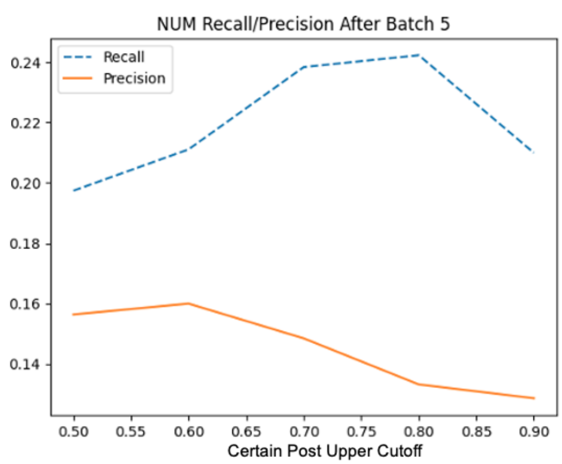


Figure 13 Model recall (dashed) and precision (solid) on synthetic data for various upper cutoffs used to define which posts are part of the set η of "certain" recruitment posts. Based on feedback from domain experts, we selected the cutoff 0.8 to optimize model recall.

Appendix C: Recruitment Post Examples

Adult Entertainment

'Elite Topless Waitresses is looking for high class models to join our amazing team. We are looking for bubbly promo girls, bikini, lingerie, topless and nude waitresses to work at private functions and exclusive events around Sydney. Experience is not necessary, training will be provided. To apply please send a couple of photos (face and body) with your name, age (must be +18) and a brief description of yourself to [redacted email] or [redacted name] [redacted phone#] If you don't have any professional photos we can organize a photoshoot with our photographer and the agency will take care of the costs. So what are you waiting for? JOIN US today!!!!'

'Having a little house party and want some quality company for us. Looking for 3-4 pretty hostesses. Must be cool, over 21, and have taste in music. Email very recent photo with subject line of a kind of car.'

'Jobs Exotic Dancers of Canada, Inc is hiring new female exotic dancers for private events and Bachelor Parties! This job allows you to make a ton of money in a very short amount of time and is perfect for college students, or anyone looking to make big weekend money for only a few hours of work! We welcome new dancers with a good attitude and the willingness to learn. Plus, we are now offering a $500 sign-on bonus to get ready for our busy season! REQUIREMENTS: 18+ Good Attitude Friendly Fit body Outgoing Personality Confident Comfortable working closely with other female dancers Reliable To apply either send us an email [redacted email] ive.comor Call us Directly [redacted phone#] Exotic Dancers of Canada). Please send an email with a full body shot (in bathing suit) and a picture of your face. Tell us what city you currently live in and a few sentences about yourself. GREAT PAY!'

Figure 14 Sample examples of recruitment posts related to Adult Entertainment

Escort Services

'(*) Very Busy OutCall Agency (*) To work with our agency, you must be at least 19 years of age (photo id required), attractive, well kept, and maintain a healthy mind, body, and lifestyle. We are a professional reputable company established in Vancouver since 1999. The safety and discretion of our clients and employees is of utmost importance, this is one of the reasons why we are the best escort agency to work with. Once hired you have the flexibility to choose your own hours and work as few or as many shifts as you desire. All escorts and entertainers must be displayed on the site to get refferals. You may provide your own professional photos, or have them done through our agency. To get an idea of how your photos will look please view the current escort gallery. We won't hire girls who wish to work for 3rd party companies, independently, or are listed on other sites. Please call [redacted phone#] Email Current Photos to Expedite the process'

'* US: Upscale, established and reputable. Featuring SAFE, 100% pre-screened professional white collar clientele exclusively. Fantasys is the oldest original owner escort agency in the entire USA for over 35 YEARS. * YOU: Upscale, SLENDER, clean and sober, stable, reliable, fit and attractive. NARCISSISTS and chubbies need not bother. * Please APPLY with accurate face and body shots to http://FantasysEscortService.com'

'A male escort (gigolo) is a man who is hired by a woman to accompany her as a companion. Some male escorts also perform more intimate favors. Male escorts may also be men who are kept by older, wealthy women over a longer period of time, exchanging companionship for money or gifts. Being a male escort sounds like a dream job, right? It IS a DREAM JOB! Just imagine being paid to entertain women, at their expense, as you live out your fantasies in a variety of exotic locations. This could be your opportunity to enter the always interesting and often exciting world of the well-paid straight male escort. You will find working as an escort to be rewarding, both financially, and also the potential introductions and connections that will come as a result of moving within these centers of influence. DO NOT HESITATE AND REGISTER TODAY! Just Few Details Of Yours And you are Good To go. We Need Following Information Over Our Whats app Height Weight Chest Biceps Body tone Body Type We guarantee you a gigolo job if you do it right NOW. We Believe In Quick Action So HURRY For Quick Acceptance. @just an registration amount of 2500Rs Only. Call Us John Martin Call/Whatsaap @ [redacted phone#]

Figure 15 Sample examples of recruitment posts related to Escort Services

Sugar Parent

'I am looking for a non pro Asian for an arrangement / sugar baby. I am looking to meet once a week. An allowance is given. Plus spoiling. Looking for more than intimate fun. Looking to go to lunch, dinners, sporting events, beach and then end our time with intimacy. I would like and hope our arrangement to be similar to an actual relationship, just minus all the stress, drama, possivesness and jealousy that can come with one. My last SB and I got to become pretty good friends. I am 46 years old, white and a successful businessman. If interested please tell me a little about yourself, if you been in an arrangement before and a picture of yourself. NOT Looking for time wasters, games or drama!'

'28 yrs old male seeking 18-29 FEMALE ONLY!sugar baby for mutual beneficial spoiling.'

"I'm 36 seeking an attractive female to be my sugarbaby must be near Alton nh or can travel send pic and where your from can do large weekly allowance"

Figure 16 Sample examples of recruitment posts related to Sugar Parents

Porn

'Currently shooting in Wilmington and are looking for females for Oral scenes. The scenes pay $300-$475 depending on duration and positions. This is 100% real so please contact if you are a female over the age of 18 and are willing to work in a professional environment with a friendly staff.'

'Be a porn star. Make a porn video. $200 [redacted phone#]

"Amateur Porn Actress(es) Needed: Hello I am searching for a woman or women 18-48 willing to star in some online adult entertainment videos. In the videos you will not be having intercourse. Nudity is a must though. Mostly masturbation. Squirting is a plus. Some videos will be scripted with a plot. Some will not. Feet play and cum on glasses are plusses but not necessary. Girl on girl a plus but not essential. Some silliness is encouraged. There will be no other men involved other than myself. I may not even get involved other than filming. But I have funny ideas where my penis is just randomly appearing places and everyone is like wtf..lol. The short films will be put onto my own online profile on a website. Pay is not immediate, but once the videos make money, the money will be distributed fairly. If I do get involved at all no intercourse or anal will be involved. Probably not even bjs. Just hjs or watching me play with myself.Videos may even be solo girl videos. Whatever you are most comfortable with. Also please have an ID or Driver's license and recent (April 2017) std check proof."

Figure 17 Sample examples of recruitment posts related to Porn

Personals

19 year old look for a europian girl around my age to just be friends with benefits at the moment, email me if you're keen.'

"55 yr old, handsome, gentleman, looking for a very SLIM, white, super clean "crack whore-type" of lover, between the ages of 18 and 68. Must have beautiful feet and enjoy having a tongue in her pussy and ass. I'm looking for the "girl next door" and not an escort or a pro. I'm available to meet on most afternoons, usually from 2 to 6 pm on most weekdays, and on Saturdays. Reply with photo of face, body and feet would be appreciated, otherwise your email will be ignored. You can also contact me at [redacted phone#] e-five-0"

'A fit 45 years young looking for a girl 18 to 30 for a (great and good) time. Please send your info and a picture so we can get together for some cafe in a public place and then we see where we can go from there. Real respond please.'

Figure 18 Sample examples of recruitment posts related to Personals

Modeling

```
'Best Kept Secret Studio is Hiring New Models world-wide -Let you guide be a team of profecionals (15y expirience) -Rise above the
Rest you get Social media training by a webmaster -90% Payouts paid 2 x a month Pay by Paxum, cosmo, Check, Direct Deposit -We offe
r over 10 Sites to work on! -We work for you! -YOUR THE BOSS! -For more info visit our site WWW.BKSSTUDIO.COM [redacted phone#]
8'
```

```
"Busy Agency Hiring!!! Do you consider yourself the best? We are currently seeking a few select models in select locations. No expe
rience necessary. All models from around the globe are welcome to apply. So, if you are: * 18-50 yrs. (Proper ID is a must). * A mo
del (runway, lingerie, bikini, fitness or nude) or just look like one. * The girl next door or college type. * Physically fit and t
oned. * Professional, easy. Make up to $1000 a night as an outcall escort. If you are attractive, sexy, have personality, and are s
erious about making good money, there is no reason to be broke! No experience OK. Must be 18 or older FULL NAME: WORKING NAME: EMAI
L: PHONE NUMBER: CITY: REAL AGE: ETHNICITY: HEIGHT: WEIGHT: BODY MEASUREMENTS: EYE COLOR: TATTOOS: PIERCINGS: ****3 PICS OR MORE OF
 YOURSELF (DOES NOT NEED TO BE PROFESSIONAL)(MUST BE RECENT AND ACCURATE)**** If any immediate questions please [redacted phone#] L
ADIES ONLY) Driver's can text [redacted phone#]
```

```
'A top online Lingerie retail store is in need of female models for a photo shoot for its December edition Interested ladies are as
ked to submit the following: Name Phone number 2 recent full length photos Email applications: [email protected] to arrange an inte
rview.'
```

Figure 19 Sample examples of recruitment posts related to Modeling

Massage

```
'Brand New SPA Hiring Beautiful Ladies (Body Rub Techs, Massage Therapists (Licensed), Phone Girls, Managers needed ASAP. Start Imm
ediately Must be 21 and over No experience necessary (Will train) Your own transportation Proper ID Flexible schedule (You work whe
n you want) Beautiful location in Dania Beach Great potential Will make a lot of money. Please for consideration send 3 of your bes
t pictures( Full, front, back, face must be seen) through text or email. You will be contacted immediately.'
```

```
'SINGLE MOTHERS -OR- FEMALES NEEDED $20-$60/HR PER JOB Come join our Massage team needing 7 more females and two male masseuse paid
 training is available $20 to $60 per hour per job if you are needing the extra income boost or financial stability look no further
!! Nine 1 zero..9threefour + thirteen27 "massage"'
```

```
'Kaya masseuse hiring attractive females for body massage. Accommodation will be provided And you will be paid weekly. Whatsapp for
 more information 1876-453-7349'
```

Figure 20 Sample examples of recruitment posts related to Massage

Appendix D: Recruitment Templates

The active learning algorithm designed uncovered more than 27 types of recruitment categories on the deep web.

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

Table 3 **Example recruitment templates identified across labeled posts**

Appendix E: Details on Synthetic Experiments

E.1. Data Generation

Recall that each post generated in our synthetic dataset will have three components: a node assignment ("origin" location), an edge assignment (e.g., connection to another location representing a recruitment-to-sales pathway) and post content. Nodes and edges are assigned probabilistically based on the volumes of activity on nodes and edges found in our TellFinder dataset within the United States and Canada in order to maintain the same underlying network structure as our original data.

Post features are generated using a Python package called "make classification" to create numerical posts that have 100 features and a binary label (e.g., recruitment versus sales) (Learn 2021). This package creates clusters along a hypercube $[0, 2]^{100}$ of the dimension of features, and assigns clusters to each class. It introduces interdependence amongst the features and adds noise. Once we have constructed the posts, we randomly assign them a label (recruitment vs. sales), then randomly assign them to a node based on the volume of sales/recruitment activity at that node, and finally randomly assign them to an edge (i.e., a target node) based on the edge volumes from that node. To ensure the post features are heterogeneous across nodes, we randomly assign a distribution shift sampled i.i.d. from Uniform($\{1, ..., 10\}$) to each node, and shift the features for all posts by the distribution shift for that node; in particular, if the post features are $\phi(x)$ and the distribution shift is $k \in \{1, ..., 10\}$, then the shifted features are $\phi'(x,k) = \phi(x) + k \cdot \vec{1}$, where $\vec{1} \in \mathbb{R}^{100}$ is the all-ones vector. We chose values for distribution shifts and noise to develop a dataset with sufficient complexity that it requires 100 batches of active learning to achieve maximum model performance. This allows us to more effectively differentiate how well each method performs (e.g., using a dataset that is too simple results in all methods performing well, making it difficult to distinguish the best approach).

Thus, the synthetic data generation process results in a set of posts that have known labels and network associations. Based on this approach, we construct a synthetic dataset with approximately 450,000 posts associated with 697 nodes.

E.2. Fair Active Learning

Here, we summarize the FAL algorithm by Anahideh et al. (2022). We consider a group fairness metric $\mathcal{F}(f)$, which compares outcomes for each group and computes disparities against a privileged group for model f. For unlabeled post x, the potential change in fairness is computed based on the expected fairness across the potential realizations of the ground truth label $f^*(x)$ for x:

$$\mathcal{F}(f;x) = \mathcal{F}(f_{x,0}) \cdot P(f^*(x) = 0) + \mathcal{F}(f_{x,1}) \cdot P(f^*(x) = 1),$$

where $f_{x,y}$ is the update of f based on example (x,y). Thus, $\mathcal{F}(f;x)$ is the expected fairness of the updated model $f_{x,y}$ if we choose to label example x. Additionally, Anahideh et al. (2022) includes Shannon entropy

$$H(x) = -P(f^*(x) = 0) \cdot \log_2 P(f^*(x) = 0) - P(f^*(x) = 1) \cdot \log_2 P(f^*(x) = 1)$$

in their prioritization metric to additionally focus on selecting points to reduce misclassification error. Together, their overall prioritization metric for post x is

$$W(x) = \lambda \cdot H(x) + (1 - \lambda) \cdot (\mathcal{F}(f) - \mathcal{F}(f,x)),$$

where the first term is the Shannon entropy for x, the second term is the expected reduction in fairness from labeling x, and λ is a hyperparameter trading off the two (chosen using 10-fold cross-validation). We take the fairness metric $\mathcal{F}$ to be the well-known notion of demographic parity (Hardt et al. 2016). This metric ensures that we also uncover recruitment activity in less populated areas.

Appendix F: Auxiliary Results

We examine a number of relevant socioeconomic indicators (summarized in Table 4) 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 (Census 2018), US Bureau of Economic Analysis (BEA 2018), US Bureau of Labor Statistics (BLS 2018), US Department of Housing and Urban Development (HUD 2019), National Center for Education Statistics (NCES 2017), WomensShelters.org (Shelters 2021), Proximity One (Proximity 2009), and US Department of Justice (FBI 2016). The data collected from these sources focuses on both economic attributes (household income, GDP, unemployment) and social attributes (homelessness, education, crime).

We run separate Kolmogorov Smirnov tests (Smirnov 1939) 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 (Benjamini and Hochberg 1995) 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 21 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.

We also examine the difference in the city sizes across recruitment categories. We find that larger cities (typically on the East and West coast) primarily had recruitment related to sugar parents, personal ads, and escort services, while mid-sized cities had recruitment related to massage and modeling, and smaller cities had activity predominantly related to adult entertainment and porn services. In Figure 22, for each recruitment type, we plot the average city populations, weighted by the percentage share (i.e., the fraction of recruitment posts of that particular recruitment type that occur in that city).

F.1. 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.

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 bachelor's 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

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$

Table 4 List of variables from governmental 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.

1. **Rubmaps:** Rubmaps.ch is a review site for massage parlors with sexual services, and has been used to assess commercial sex activity in prior work (Bouche and Crotty 2018, Diaz and Panangadan 2020). 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.

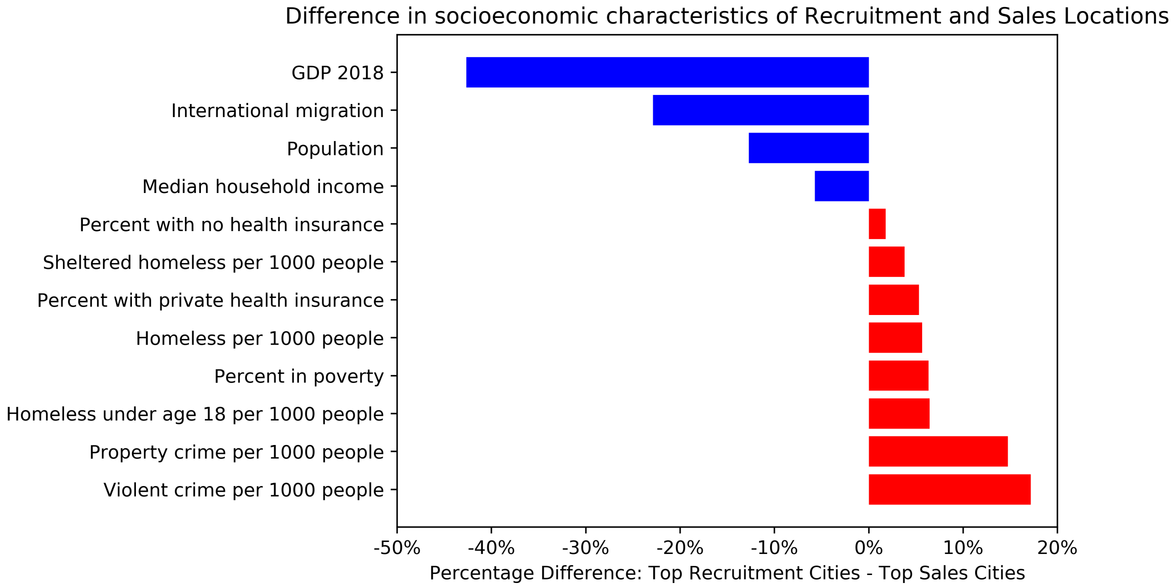


Figure 21 **Comparing selected socioeconomic variables for the top 50 sender (recruitment) and receiver (sales) cities in the United States. Blue and red bars indicate variables with higher values in receiver and sender cities respectively.**

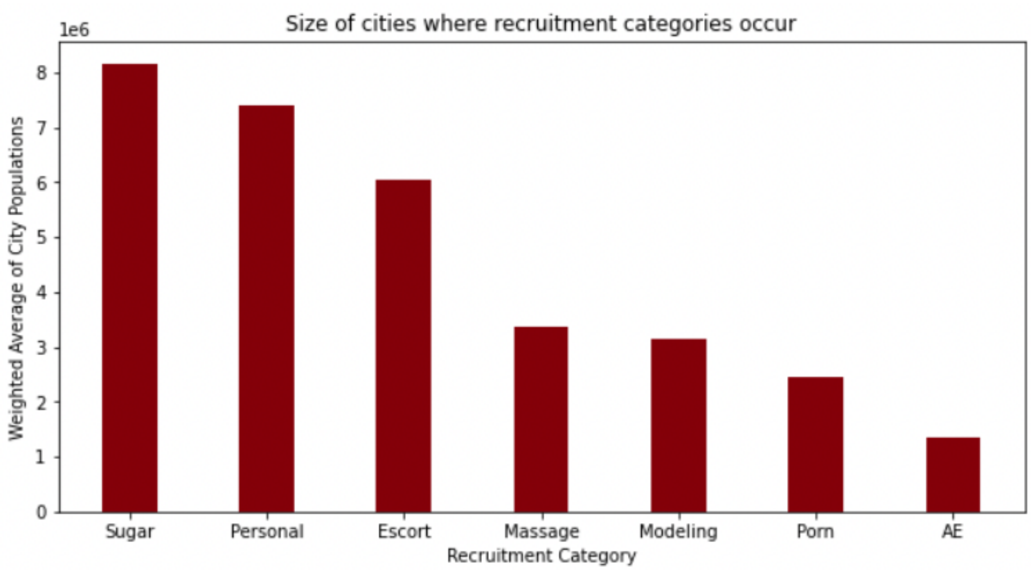


Figure 22 **Comparison of the weighted population counts for each recruitment type. Larger cities focus on recruiting related to sugar parents, personal ads, and escort services; smaller cities predominantly recruit for adult entertainment and porn.**

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 com-

mercial 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 23). 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 23 **Empirical distribution of commercial sex activity in the United States inferred from Google Places, Rubmaps, and Deep Web respectively**