AFRL-RI-RS-TR-2018-230

MACHINE LEARNING FOR ADAPTABLE HETEROGENEOUS INDEXING AND SEARCH

CARNEGIE MELLON UNIVERSITY

SEPTEMBER 2018

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2018-230 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
EDWARD DEPALMA
Work Unit Manager

/ S /
TIMOTHY A. FARRELL
Deputy Chief, Information Intelligence
Systems and Analysis Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
SEPTEMBER 2018	FINAL TECHNICAL REPORT	Sep 2014 – Mar 2018

4. TITLE AND SUBTITLE

MACHINE LEARNING FOR ADAPTABLE HETEROGENEOUS INDEXING AND SEARCH

5a. CONTRACT NUMBER
FA8750-14-2-0244

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62702E

6. AUTHOR(S)

Artur Dubrawski

5d. PROJECT NUMBER
MEMX

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2018-230

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This work has developed new technology that matches content between imagery relevant to a certain domain. The focus of this effort was on a law enforcement application concentrating on Human Trafficking (HT) in order to identify the room and background images that would appear in illicit advertisements and connect them to those available on hotel websites. These images would then be gathered according to hotel chain, and then further refined to individual location which would lead to estimates of GPS coordinates. Contractor also employed their unique facial recognition methods to match individuals across ads. The technology is being validated through transition to multiple law enforcement agencies.

15. SUBJECT TERMS

Human Trafficking (HT), Facial Recognition, Hashing, Facial Analytics, Image Matching, Machine Learning, Marinus Analytics, MEMEX, Cybercrime

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			EDWARD DEPALMA
U	U	U	UU	20	19b. TELEPHONE NUMBER (*Include area code*) 315-330-3037

CONTENTS

LIST OF FIGURES

1 SUMMARY

The challenges to performing investigative efforts on data from the deep web include: content with multiple, heterogeneous modalities (e.g. text, images, video, audio); dynamic content whose temporal patterns contain important information; and investigators' goals that are inherently adaptive. We developed statistical machine learning methods to address these challenges, and made them available to the MEMEX community and beyond. The main components of the effort were:

Image. We created tools for matching content between images. Specifically, we targeted whole image, pose, and background based similarities. We developed and made available an image similarity search capability based on the cosine metric computed using the inner layers of an existing convolutional deep network. That work was extended to include deep hashing, scalable ITQ hashing, and task relevant image similarity tools. This capability will help track illicit activity. We also deployed our face recognition methods to match individuals across escort ads.

Text/Language. We developed information extractors and evaluated transformations on unstructured text to enable classification and modeling of complex textual objects such as escort advertisements.

Multi-modal Aggregation/Tracking. We developed a proclivity measure with desirable theoretical properties for analyzing attributed networks. We made contributions to the entity resolution sub-domain in machine learning, publishing tools and theoretical properties that focus on the problem of scaling entity resolution to very large data sets. We demonstrated the value of analyzing aggregate data for understanding temporal trends and identifying anomalies. Finally, we developed a multi-modal active learning algorithm of learning important domain/task-specific relationships in complex data.

Integrated System. Marinus Analytics, a CMU Auton Lab anti-trafficking spinoff, performed use case development, user testing, and deployment of technology we developed through its law enforcement partners. These efforts are most saliently demonstrated through multiple successful recoveries of missing minors. For example, in one case, our image similarity search led to the recovery of a runaway child from foster care.

Overall we deployed over 10 technical components, most on github, making our work available to the MEMEX community and beyond. CMU affiliates gave 5 invited lectures and tutorials on directly relevant topics, and Marinus Analytics gave 12 program related talks. We received an innovation award for data science and published 10 peer-reviewed articles. Our work involved and contributed to the education of over 35 students.

2 INTRODUCTION AND KEY ACCOMPLISHMENTS

Domain specific search requires methods that can be adapted, either explicitly or implicitly, to a novel problem space with minimal effort. Some problems in search are generally ubiquitous, such as retrieval of documents/images, bearing strong similarity. However, for some applications the notion of related or similar might be unique and highly domain-specific. To address this difficult challenge we identified three main goals, which are itemized below along with the related key accomplishments.

Develop new algorithms. We advanced the state of the art for algorithms in the areas of entity resolution, annotated graph analysis, and image similarity search. In computer vision, we improved the ability to identify related images by methods for supervised hash code generation, automatic background detection, and scalable hash-based search. In entity resolution, we improved understanding of the challenges

inherent in large-scale data sets and provided tools to overcome them. In predictive modeling we developed tools for overcoming biased performance measures. In domain-specific search, we developed an active graph construction algorithm for interactively learning to recognize important relationships and content.

Develop an integrated software tool. We made our algorithmic contributions available to the MEMEX community of teams constructing end-to-end systems. Libraries containing these components are available in github repositories described in Section 4.12. Our image similarity search capabilities were integrated into the Traffic Jam tool.

Field the software tool. Through our law enforcement partners, we made our research results available, resulting in several successful recoveries of minors and incarceration of multiple perpetrators.

3 METHODS

We developed a wide variety of tools and methods to address the challenges of domain-specific search. Our image-based tools rely on convolutional neural networks, and advancement of technology was largely gained by problem formulation and clever use of the constituent elements of the technology (e.g. internal layer activations as vector space representations). Other non-image based tools largely rely on statistical modeling and analysis.

4 ASSUMPTIONS AND PROCEDURES

Program goals were largely meant to be application agnostic; the development of tools for domain-specific search. However, we found it useful to focus our efforts on a specific application, namely counter human-trafficking (CHT), as it inspired many interesting approaches and revealed numerous pitfalls that might not have been fully appreciated were one to work only in the abstract. We feel that many of the problems in the CHT space, such as search for elements representing a specific entity or related entities, can be generalized to other applications. For this reason, most of the project activities and results below are presented in a CHT application specific way. Our team has also demonstrated some of our developed technology on the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) applications. We have also effectively dabbled in the countering micro-cap fraud application.

5 SUMMARY OF RESULTS AND DISCUSSION

Below we describe a selection of project activities, results, and findings. In general, we developed many new algorithms and methods for addressing challenges in real-world machine learning problems and extending the utility of existing approaches. Section 6 enumerates publications funded or partially funded by this project.

5.1 Escort ad classification

The digital presence of prostitution and potential sex trafficking activities provide a rich, accessible, and affordable source of information. This information is readily available in the public space and is a formidable tool for quantifying prevalence, characterizing the involved populations, observing their variation over different strata, as well as dynamic changes of the market activity and operating principles over time and space. It can also be used in the practice of law enforcement to identify new leads, monitor suspicious activity, support building cases for prosecution, and enhance victim-oriented policing. It can help social workers and community support organizations to identify, track, and rescue victims of sex trafficking. At the outset of the MEMEX program, these data were underutilized partially due to their sheer abundance that makes manual searches for evidence extremely laborious and, consequently, spotty, and partially due to lack of appropriate tools that could be put in the hands of those in need in order to unleash the power of data in all of the above mentioned application scenarios.

To help remedy these issues, we attempted to train machine learning classifiers that could distinguish ads that represented cases of human trafficking from relatively benign cases of 'garden variety' prostitution. Initially we used several different feature sets to attack the classification problem. Figure 1 shows the

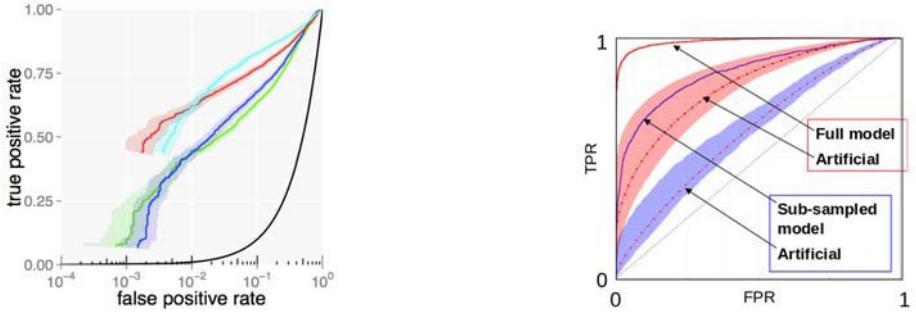

Figure 1: ROC curves for escort ad classifiers using different featurizations (left), and using random bootstrapping to reproduce the inherent evaluation bias (right).

ROC curves for four random forest models trained using domain expert identified keywords (green), automated information extractor results (blue), bag-of-words (cyan), and bag-of-words with the information extractor results appended (red). In cross validation, these classifiers all perform quite well. However, this is largely due to information leakage as described in Section 4.6.1. We attempted to correct for this difficulty by reproducing the dependency structure of the labeled data (collecting all ads with related phone numbers) in random bootstrap experiments, see Figure 1. This experiment demonstrated that indeed the practice of labeling ads by their constituent phone numbers led to a significant positive bias in evaluation and, further, that the bias did not fully go away simply by blocking the cross-validation routine on phone number. This fact was brought home to the other project teams during the spring 2016 QPR, in which many teams trained similar classifiers. Our team assumed a clique dependency structure based on phone number and sampled a single ad from each clique to train the classifier. As a result, our team's classifier produced good performance on the hold out set, whereas other teams found that their models gave near random performance (despite strong initial evaluation using cross-validation).

In other related activities, we produced a logistic regression model based on a bag-of-words featurization that highlighted the words that were critical to the final classification result, thus explaining the model's decisions. This model was also made available to the MEMEX community.

5.1.1 Escort ad featurization and information extraction

For the purposes of analysis, it was essential that individual escort advertisements be featurized. We developed a regular expression based information extractor for escort ads [6], which identified 14 types of information and extracted their values from noisy text. There are several approaches one can take to information extraction from text, varying in complexit y, strengths, and weaknesses. Perhaps the simplest approach is to compile a dictionary of terms of interest and identify occurrences of these key-words in the ad texts. The hope is that either sheer appearance or frequency of appearance of some of these terms in an ad, or their combinations, would be useful in sorting apart advertisements of specific interest from the rest of data. Our key-word dictionary was compiled through several interviews with law enforcement investigators and as such tuned through their experience to identify ads that are operationally suspicious and may be indicative of trafficking. We compiled 115 such key-words and short phrases.

Alternatively, one might train statistical models, such as e.g. conditional random fields, to identify the textual elements of interest, or apply information extraction techniques such as Named Entity Recognit ion and Classification. Informal, noisy, partially obfuscated text prevalent in escort ads requires however a substantial amount of formally annotated training data for those modern approaches to succeed. Such data is often costly and laborious to obtain, especially given that based on our experience, the vocabulary of advertisements actually evolves overtime.

For some tasks such as clustering or classification of advertisements it is often desirable to rely on low dimensional, fixed length numerical feature vector representation that relaxes the requirement that the individual feature values are readily interpretable. We experimented with domain-expert-recommended key-words, with regular expressions over annotations based information extraction methodology, and we also featurized sex advertisements data using unsupervised bag-of-words representation compressed using PCA.

Our regular expressions approach was constructed to extract ostensibly personal identifying physical and operational characteristics. We focused on domain specific attributes that may be useful in characterizing individuals or groups responsible for each advertisement. We used the General Architecture for Text Engineering (GATE) software framework to construct a hierarchy of rule sets tuned to extract key pieces of data that may be informative of trafficking from ad text. Briefly, these rule sets first identified numbers and then phone numbers, before passing the text through a battery of rules constructed to identify specific features of text. This extractor identified age, cost(s), email, ethnicity, eye color, hair color, name, phone number, restrictions, skin color, url, height, physical measurements, and weight. Many of these features are self-explanatory. “Restriction” refers to common limitations of the johns an escort is willing to work with such as “no African Americans” or “no men under 30.” “Url” refers to web addresses that are explicitly indicated in the text, not hyperlinks. “Physical measurements” refers to any indications of cup, chest, waist, and/or hip sizes. The set of rules was constructed by hand using a random selection of 1,000 advertisements for female escorts and evaluated on a separate random selection of the same size.

5.2 Entity resolution

In [8], we employed match-and-merge entity resolution techniques to aggregate escort ads for the purposes of identifying cases of human trafficking. It was thought that attacking the classification task at the aggregate level would provide a much richer characterization of the underlying behavior and thus provide a strong signal of nefarious activity. The approach used ‘strong’ links between ads (sharing a phone number) as a proxy ground truth for the relatedness of ads. A random forest was then trained to identify ads that ought to be related without the benefit of using the phone number. The strongest predictor was the amount of overlapping text. This was not surprising as there was a significant amount of near duplicate ads in the data set. The entity resolution was difficult however due to issues with snowballs (accumulation of the majority of observations into a single entity) and computational cost.

Both of these issues were overcome to some degree by blocking the sets of ads that were compared by comparing only those ads that share rare unigrams or bigrams. Resolution was carried out within blocks first, then across blocks. While the entity resolution scheme was successful overall, it led to only a small improvement in human trafficking classification. However rule learning showed promise as an interpretable method for learning how to identify cases oftrafficking. Match-and-merge is not the only approach to entity resolution. The planted partition model is another common approach. In [2], we extend this approach to consider features on the edges of a graph. Here we imagine our data (e.g. ads) are related through connections (common phone number, name, etc.) and that these connections are described by features. Our approach was to learn the distribution of edge features given a pair of connected or disconnected observations. From these probability distributions we recover the most likely partition of the observations. The key insight from our approach is that multidimensional edge features can be used to effectively learn structure in clusters. Relationships in real world data are more complex than a simple scalar similarity function, and our methods can benefit from capturing that additional complexit y. Then we can use the learned cluster structure to both determine the correct number of clusters and to handle situations where we are given new, previously unseen clusters, by assuming similar structure.

One significant challenge to scaling entity resolution algorithms to massive datasets is the dynamics involved in performance changes moving from small training/evaluation datasets to the large datasets encountered during application/deployment. Unlike traditional machine learning tasks, when an entity resolution algorithm performs well on small holdout datasets, there is no guarantee this performance holds on larger hold-out datasets. To characterize this phenomenon, we proved simple bounding properties between the performance of a match function on a small validation set and the performance of a pairwise entity resolution algorithm on arbitrarily sized datasets [3]. These bounds also enable optimization of pairwise entity resolution algorithms for large datasets, using a small set of labeled data. This was demonstrated on 4 datasets including 10K escort ads.

5.3 Aggregate level analysis

In [6] we demonstrated how the multivariate temporal scan algorithm (a part of our Temporal Anomaly Detection, TAD, tool) could be used to identify unusual changes in posting volume. We used phone numbers to connect ads and produce a behavioral stratification: local, new-to-town, and first-appearance. The new-to-town ads represented ads that were related to prior posting activity, but not the ad's current location. First-appearance ads were those not related to any prior activity. The remainder were local. By analyzing the new-to-town ads, we were able to uncover interesting trends, such as an influx of sex-workers to North Dakota coinciding with the oil boom, and quantitatively investigate the relationship between sex-worker activity and large scale public events (such as the Super Bowl).

In [4], we conducted a comprehensive analysis of advertisement trends over all of our data by identifying spikes in sex ad activity and by the assumed implication of our proxy measure, spikes in sex trafficking activity. We reason that if our proxy measure does indeed capture responses of the online escort market to large public events, we should be able to uncover anomalous exceedances of such activity that co-occur with such events. We then considered a broad sample of public events, including but not limited to recent Super Bowls, to study anomalies in an attempt to quantify evidence of their potential impact on trafficking activit y. We examined 38 different public events, chosen for attendance numbers and duration comparable to the Super Bowl from a diverse range of types such as sporting events, festivals, and conventions. Figure 2 shows a bubble plot of the exceedances associated with different events. It also shows the time series of new-to-town ad counts for Myrtle Beach, SC, one of the most significant yet unexpected patterns uncovered by the temporal scan algorithm. Our analysis puts the impact of local public events on sex advertisement activity into perspective. We find that many of the events we considered at random are not correlated with a statistically significant impact on sex-worker advertising,

while evidence of others that are correlated was uncovered by our comprehensive analysis over all data. Details can be found in the recently published paper [4]. Under the vein of multi-modal characterization of observation sets, we applied canonical autocorrelation analysis to model groups of escort advertisements associated with human trafficking [5]. This is a descriptive analytic task, wherein the objective is to extract and quantify important patterns that are difficult for humans to find. Such characterizations are important to law enforcement agents who track and apprehend human traffickers. We hope that tools like this will deepen their understanding of how each network operates including who are their victims, who are their clients, and where and how they operate.

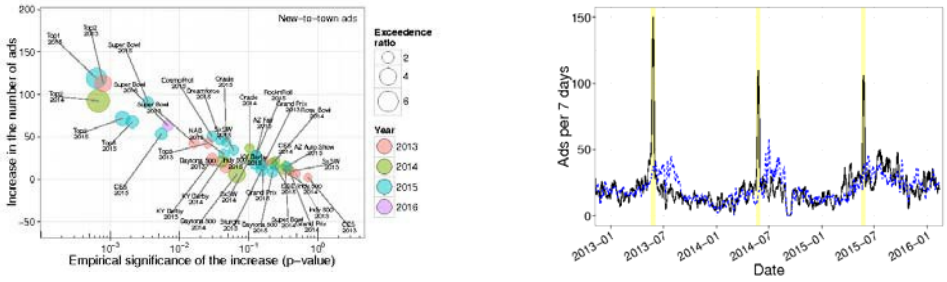

Figure 2: Bubble chart (left) showing the significance of ad count exceedance for various large scale public events, and the time series of new-to-town counts for the most significant event found in the data (right).

This effort was undertaken using two machine learning techniques: Canonical Correlation Analysis (CCA) and Supervised Latent Dirichlet Allocation (sLDA). The sparse-by-design output of both methods are intuitive to interpret for non-machine learning experts, making them particularly suitable for projects involving law enforcement. While the variant of CCA we developed, Canonical Autocorrelation Analysis (CAA), characterizes relationships amongst features in each trafficking ring, sLDA discovers topics which describe how features relate across groups of trafficking rings.

Using 13M ads, we used the regular expression based information extractor described above to generate 34 features that included physical descriptions and location information. Ads were aggregated by phone number and those aggregates that contain one of 1,700 phone numbers known to be associated with cases of trafficking were identified as 'positive.' CAA successfully uncovered related clusters of ads by uncovering patterns of phone number area code and posting location (which sometimes differ), though sparseness and discrete nature of features made CAA a bit difficult to wield in this application. sLDA found topics which described behaviors across various human trafficking clusters. Human trafficking ads focus more on ethnicity than the negative topics, especially exotic ethnicities such as Thai, Persian, and Cuban. They are also likely to exaggerate the physical features of women advertised. Additionally, human trafficking ads are unlikely to disclose the ages of women in advertisements. These characterizations are natural for human traffickers who tend to traffic women from foreign countries, many of whom may be minors.

5.4 Multimodal match functions for HT applications

Many real world data sets can be characterized as attributed networks. In social networks nodes usually represent people and edges friendships. However, social sites often capture additional attributes such as interests or demographics of individuals. In the CHT application, escort ads contain many attributes (e.g., pricing, physical characteristics, etc.) and links (e.g., phone numbers). However, despite the prevalence of attributed graphs, the vast majority of network science has focused so far on leveraging graph structure/topology, ignoring the attributes. It has been noted that similar individuals tend to be associated, and associated individuals tend to be similar. This assortative mixing and peer influence results in a

homophily pattern observed in many real world networks, where neighboring nodes exhibit similar characteristics/attributes. When assortativity is used as a measure for structural correlation of a single attribute it suffers from some significant shortcomings, Specifically, assortativity (r-index) can identify perfect homophily but it is unable to distinguish between perfect heterophily and randomness. Further, it cannot characterize or distinguish the mixing patterns involving pairs of attributes. We addressed the shortcomings of the current state-of-art by defining a formal characterization of proclivity [9] in attributed networks: the inclination or predisposition of nodes with a certain value for an attribute to connect to nodes with a certain other value for the same (self-proclivity) or a different attribute (cross-proclivity).

Our proclivity measure is constructed as the divergence of a confusion matrix of a pair of attributes. The elements of the matrix represent the number of edges that link two nodes bearing the associated pair of values. We have proven several theoretical properties of the measure, demonstrating that it is superior to the alternatives in consistency and generality, without sacrificing scalability. The proclivity measure quantifies the strength of the relationship between attributes in terms of how well one can predict the other, subject to the local graphtopology.

5.5 User-specific query adaptation

We developed a first approach for modifying search algorithm behavior to suit user specific needs. In many cases search infrastructure is fixed, e.g. hash tables are precomputed and expensive to update. In these cases, one may still alter the behavior and performance of a search engine by modifying the query. We considered a region of interest (ROI) scenario in which a user is interested in searching for contents similar to a seed document, but only for a subset of seed. In this case, it is most reasonable to view the query not as a single seed document but rather a distribution over seed documents. I.e. we take the query to be the distribution of all documents that contain the region of interest. While this distribution may not be known explicitly, in many cases it is possible to sample from it. In the case of images we sampled images, pasted the ROI on them, pushed these 'Frankenstein' images through the search engine's featurization and averaged the results. This average vector was the query vector. For many common choices of similarity function (e.g., inner product, cosine similarity, squared Mahalanobis distance, etc.) finding the document that maximizes expected similarity (over the implied query distribution) is equivalent to maximizing similarity with the average query vector. In these scenarios, one can submit the mean query vector to the search engine to provide user-specific search capabilities.

5.6 Dealing with noisy real-world data

An all too common, yet under-appreciated, problem in real world applications is dependency and information leakage in curated data sets. This issue is demonstrates itself well in the pernicious example of trafficking victim classification from escort advertisements. A high degree of difficulty associated with of acquiring ground truth examples of ads representing victims of trafficking resulted in relatively modest (in modern machine learning standards) labeled data set sizes. Moreover, well intentioned efforts include all/many ads associated with a single case. On top of that, many cases may be interrelated. The consequence of this is that the adjudicated data with which one might train models for domain specific search tasks contain large degrees of dependency across observations.

5.6.1 Biased performance estimation

When dependency exists in labeled data sets, model evaluation can be biased if the dependency is not accounted for. In [1], we show that while controlling for known dependency structure by appropriately structuring folds used in cross-validation is nearly unbiased, it is insufficient if the dependency structure is imperfectly known. For example, in the CHT application, near duplicates and related ads may be somewhat difficult to identify with complete certainty. Creating cross-validation folds based on imperfect understanding of the underlying dependencies results in information leakage from the testing data making its way into training in a manner that cannot yield good generalization in practice. The result is overly optimistic estimation of model performance.

To address the issue of information leakage in cross-validation we developed a binomial block bootstrap estimator of the true generalization performance of the model [1]. The approach is to intentionally inject additional information leakage in the form of adding test or training set observations to the opposing set. By measuring performance under additional corruptions, one can extrapolate back to the uncorrupted state.

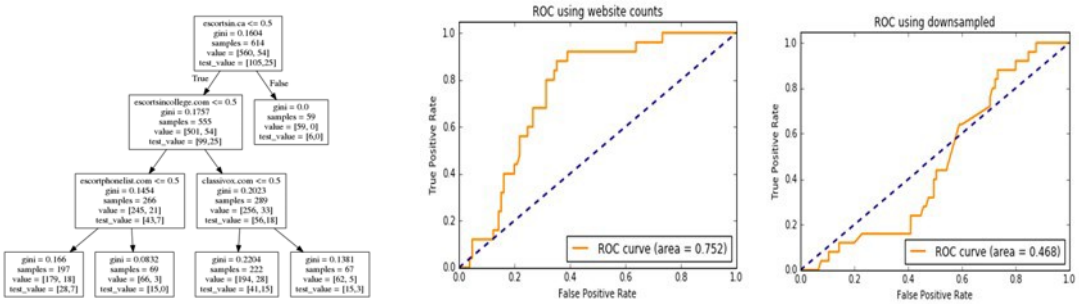

Figure 3: Graphical depiction of systematic information leakage in escort ad collection process. The decision tree on the left shows how ads can be discriminated using url. The ROC in the center shows how groups of ads can be effectively classified using only counts of url domains. Finally, the right ROC shows performance after cleaning.

5.6.2 De-biasing data with Explainable ML

Upon conducting an analysis of feature grouping in curated data, we discovered that positive and negative clusters could easily be split by url domains (for example see the simple decision tree in Figure 3). The indicates information leakage stemming from the data collection strategy. The leaked information can potentially leave some footprint in text or image contents of advertisements, and plague both training and testing data. The consequence of ignoring these issues is that supervised models may pick up on these structural regularities which, while informative for the training data, have no reason to generalize well. The ROCs in Figure 3 illustrate the scale of the problem. The first ROC shows the predictive power of knowing only which websites a cluster of ads came from. The second shows model performance after removing near duplicate negatives. The naïve approach will result in overly optimistic model evaluations.

5.7 Active graph construction

We developed active learning methods for mining attributed networks for meaningful pieces of information [10]. Using the CHT application as an inspiration, we developed a model that learns the importance of feature categories as well as individual feature values in determining the relevance of unlabeled observations. The approach operates on a k-modal evidence graph wherein nodes are assigned one of k+1 types. The principle type are entity nodes (e.g. escort ads), and the remaining k types are referred to as evidence nodes and consist of attributes or constituent elements of the principle nodes. Links between principle nodes and evidence nodes represent `has-a' or `is-a' types of relationships. The

algorithm's goal is to efficiently gather relevant principle nodes, given an initial seed. However, it tries to learn how to identify relevance on-the-fly from user behavior. It does this by assigning an importance weight to each modality. The weight describes how significant it is if two principle nodes share a value of the associated modality. Principle nodes are then ranked by their total connectivity to an iteratively constructed `in-group' initialized with a seed node (note that the seed nodes can be provided by ad classification combined with entity resolution algorithms, as in Traffic Jam tool). Weights are updated as principle nodes are added or passed over based on the user feedback received through the active learning process, and their states can be initialized from prior experience. This approach has shown promise as a first step to interactive AI for case building, one of the key remaining bottlenecks in the CHT investigative/prosecutorial practice.

5.8 Image analytics

Our team developed a number of useful image-based analytic tools. Initially we developed and made available an image similarity search capability based on cosine similarity computed using the inner layers of an existing convolutional deep network. That work was subsequently extended to include deep hashing, scalable ITQ hashing, task relevant image similarity tools.

5.8.1 Deep Hashing

We developed deep hashing methods using triplet supervision [11] for better image similarity search. Hashing is one of the most popular and powerful approximate nearest neighbor search techniques for large-scale image retrieval. Most traditional hashing methods first represent images as off-the-shelf visual features and then produce hashing codes in a separate stage. However, off-the-shelf visual features may not be optimally compatible with the hash code learning procedure, which may result in sub-optimal hash codes. Recently, deep hashing methods have been proposed to simultaneously learn image features and hash codes using deep neural networks and have shown superior performance over traditional hashing methods. Most deep hashing methods are given supervised information in the form of pairwise labels or triplet labels. The current state-of-the-art deep hashing method DPSH, which is based on pairwise labels, performs image feature learning and hash code learning simultaneously by maximizing the likelihood of pairwise similarities. Inspired by DPSH, we developed a triplet label based deep hashing method which aims to maximize the likelihood of the given triplet labels. Experimental results show that our method outperforms all the baselines on CIFAR-10 and NUS-WIDE datasets, including the state-of-the-art method DPSH and all the previous triplet label based deep hashing methods. Our method shares weights in the convolutional layers, prior to producing a hash code from fully connected layers for each of the three inputs (query, similar image, dissimilar image). The deep network is then trained to simultaneously learn the weights of the convolutional layer and to produce hash codes which match the query and similar image and not the query and dissimilar image.

5.8.2 Task-relevant features for image matching

In domain specific search, images represent a particularly challenging data type. Since images are generally compositions of many objects, meaningful comparison can depend heavily on the context. Thus, we developed domain specific (task-relevant) learning strategies for featurizing images in such a way as to provide the best similarity comparisons [7]. Our approach can be interpreted as a type of feature reweighting. The basic idea is that by looking at data, one can determine that certain parts of an image are not likely to be informative and can thus be down-weighted or suppressed entirely.

To demonstrate the approach, we used the location an image was taken as the ground truth for relatedness. If two images were taken in the same city they were considered related. An image patch is then considered useful if it is able to retrieve images taken in the same location. We then trained a model to

predict this informativeness from novel image patches. The predicted informativeness is aggregated over all image patches to produce the final reweighting for an image. Our approach outperforms existing state of the art by more than 10% mAP. It also improves the state of the art on the Oxford5K dataset when not using the provided query boxes.

5.8.3 Face analytics

We integrated our newest face detection tool into a web API. Instructions of use and a simple client example were made available on the MEMEX wiki. On average, an image takes between 200 to 400 ms to process an image using a POST request on the server. This estimate does not include the time to download the image, which depends on where the image is being downloaded from. Additionally, we developed new face matching algorithms for computing similarity scores between pairs of faces or for finding a specific person in a corpus of data.

5.9 HT domain expertize transfer

Our subcontractor, Marinus Analytics, possesses subject matter expertise of the online commercial sex industry and uses this information in proactive sex trafficking investigations by law enforcement. These strengths contributed to the MEMEX program in a number of crucial ways.

Ground –Truth & Hack-a-thons: We leveraged our wide-spread network of law enforcement contacts to capture ground truth for recurring hackathon CP1 challenges, to train computers to identify possible sex trafficking victims amongst the postings of at-will providers. We collected over 600 phone numbers tied to human trafficking investigations and cases which were permitted to be used for research purposes. We also disseminated “feedback from the field” to share investigative stories and specific attributes of these cases with the researchers and academics across MEMEX teams to enhance their understanding of the needs of law enforcement and opportunities for proactive policing. Our subject matter expertise and thought leadership result from our length and depth of interactions in this area; to-date, we have worked alongside the law enforcement community for 6 years. We regularly attend and present at major conferences relating to modernized policing, cybercrime, and human trafficking. We have shared our subject matter expertise with a wide range of stakeholders including Attorneys General, ICAC (Internet Crimes Against Children) Task Forces, lawmakers, as well as corporate meetings and international convenings such as the United Nations and the International Chiefs of Police (a list of highlights of speaking engagements and participation in events is listed at the end of this report). This key understanding is irreplaceable to the fight against human trafficking to 1) maintain a pulse on the ongoing evolution of the problem across the spectrum of communities facing this problem, 2) provide thought leadership to guide future innovation and proactive responses to criminal behavior, and 3) most importantly, to ensure resources applied to research and development deliver solutions that have measurable results, are easy to use, and make a tangible difference in real cases on the ground.

Collaboration: In addition, we collaborated with different teams through the course of the program to assist with screening of results, provide in-depth feedback, and give hands-on support of intelligence-gathering and sting operations, such as those during the Super Bowl events.

We regularly provide intelligence and operational support to law enforcement during local and federal operations including stings, Operation Cross Country, and the 3 most recent Super Bowls, including notably, Super Bowl LI in Houston, Texas this year. Marinus Analytics had a presence on the ground in Houston, supporting targeting, intelligence-gathering, and investigations. Marinus collaborated with

Carnegie Mellon's team to apply their algorithms for generating, vetting, and distributing suspicious leads with investigators.

As a result of the most recent Super Bowl operation:

5.9.1 44 human trafficking targets were generated,
5.9.2 2 national networks were identified,
5.9.3 8 target case packages were pushed to Law Enforcement within a 72-hour period,
5.9.4 Using this intel, Houston law enforcement made arrests of at least one group including a trafficker, and achieved a rescue of a sex trafficking victim.

Most recently, we worked with MIT Lincoln Laboratory (MITLL) to apply their authorship identification tools to Marinus Analytics data. Over the years, we have observed HT advertisements involving misspellings or unusual phrases with broken English, which appear concurrently in multiple cities. This suggests the presence of an organized criminal enterprise of a geographic spread. Given the difficulty to effectively police across fragmented jurisdictions, our aim is to provide clarity into the extent of these criminal operations to support a federal law enforcement response. For our experiments, we have documented an extensive ground truth collection effort, specifically tied to Asia transnational rings, and are strategically exploring text hashing and authorship identification algorithms to help uncover the extent of these organized crime networks involved in trafficking.

We have also assisted in supporting MEMEX multimedia group for targeting specific experiments, providing image sets, and providing feedback on results.

Deployment of MEMEX Tools: We have gathered our team of computer scientists, who work to enhance the open source intelligence service for law enforcement available through the Traffic Jam platform. This has included deploying MEMEX tools such as image similarity search which has been incredibly useful in identifying evidence on cases. We also leveraged the ground truth and lessons learned from CP1 challenges to improve our system alerts which raise suspicious ads to the attention of law enforcement. This enables law enforcement to be proactive—and not just reactive—in their lead generation and case building, saving them crucial time and helping them identify new leads they would not otherwise have found.

MEMEX Surrogate User Group: Another way we provided subject matter expertise was through the Surrogate User Group (SUG). We worked with other members of the SUG to test out beta versions of new MEMEX tools and research products, and gave detailed feedback from the perspective of not only a prospective user, but from the view of a law enforcement user. We held collaboration meetings with SUG members to test and review tools, and discuss our observations; then we worked with the team to document all our feedback through a report. This ensured that our feedback reached the research teams, and fostered development of new tools that would meet the specific needs and requirements of law enforcement users, fighting human trafficking as well as other crimes.

5.10 Assessment of feasibility of de-identification of MEMEX HT data

We have analyzed the challenges and risks involved in the potential publication of the MEMEX HT data repository. Our findings and recommendations have been summarized in a report titled "Anonymization of Crawled Escort Ad Data" that has been provided to the MEMEX Program Management office.

5.11 Dissemination

Our work has been distributed in top machine learning conferences, including KDD and UAI, and peer-reviewed journals. Section 6 enumerates publications funded or partially funded by this project. This project was represented in 17 talks given by our team, 1 conference award, 13 publications. This project has supported over 35 students and 2 post-docs.

5.11.1 Invited lectures, plenary talks and tutorials by the CMU team

1. Impactful Applications of Machine Learning. Warsaw University of Technology, Department of Mechatronics, Warsaw, Poland, January 3rd 2018.
2. Impactful Applications of Machine Learning. CMU ISA Executive Short Course on AI. Washington, DC, July 31st 2017.
3. Machine Learning for Societal Impact, Warsaw University of Life Sciences, Warsaw, Poland, January 2nd 2017 (invited).
4. Machine Learning and its Impactful Applications. EQT Idea Showcase Symposium, Pittsburgh, PA, November 7th, 2016 (invited).
5. Societally Beneficial Applications of Artificial Intelligence. Warsaw University of Technology, Department of Mechatronics, Warsaw October 8th 2015.

5.11.2 Presentations by Marinus Analytics

1. Crimes Against Children Task Force Meeting, U.S. Attorney's Office for the Western District of Pennsylvania, September 2017, Pittsburgh, PA.
2. Thomson Reuter's Anti-Slavery Summit, "Artificial Intelligence to Fight Human Trafficking," August 2017, Hong Kong.
3. "Facial Recognition to Rescue Sex Trafficking Victims," U.S. Attorney's Office for the Western District of Pennsylvania, July 2017, Pittsburgh, PA.
4. "Unlocking Public Data to Fight Trafficking," National Cyber Crime Conference, April 2017, Boston, MA.
5. 2017 National Cyber Forensics Training Alliance (NCFTA) Forum AHT Panel, May 2017, Pittsburgh, PA.
6. United Nations Call to Action for Gender Equality, "Marinus Analytics's work fighting human trafficking with online data," May 2016, New York City, NY.
7. Human Exploitation and Trafficking Institute Blue Ribbon Commission, Testimony on the Technology Panel," March 2016, Sacramento, CA.
8. Government Transformation Conference, "Big data to inform sex trafficking responses and improve cross-sector collaboration," February 2016, Sacramento, CA.
9. Greater Pittsburgh Nonprofit Partnership Summit, "Human Trafficking and Non-profit/Private Collaboration," October 2015, Pittsburgh, PA.
10. Conference of Western States Attorney General, "Technology to Combat Transnational Crime: Human Trafficking," July 2014, Park City, UT.
11. International Chiefs of Police Conference, "Unlocking Publicly Available Evidence to Fight Human Trafficking," October 2014, Orlando, FL.
12. Internet Crimes Against Children Task Force Commander's Annual Meeting, "Building Sex Trafficking Investigations with Traffic Jam," October 2014, Washington, DC.

5.11.3 Conference awards

De Arteaga M, Dubrawski A. "Discovery of complex anomalous patterns of sexual violence in El Salvador", Data for Policy 2016, London, UK, September 2016. 1st place Innovation Award on Data Science.

5.12 Software

We deployed a number of capabilities to the MEMEX community and beyond. Below is a list of our deployments.

1. Image Search - Feature Extraction: https://github.com/Minione/ComputeFeatures
2. Image Search - Service Deploy: https://github.com/Minione/ScalableLSH
3. Deep Hashing (DTSH): https://github.com/Minione/DTSH
4. BigITQ: https://github.com/Minione/BigITQ
5. Text based clustering tool using KwikCluster & MinHash: https://github.com/mbarnes1/kwikcluster
6. Temporal Anomaly Detector (TAD): https://github.com/autonlab/tad
7. TJBatchExtractor: https://github.com/autoncompute/CMU_memex
8. Ultron Face Detection API: https://memexproxy.com/wiki/display/MPM/CMU+Face+Detection+Service
9. Active learning: https://github.com/benbo/MAL
10. Escort ad classifier: (private due to sensitive nature) https://memexproxy.com/wiki/display/MEM/Carnegie+Mellon+University#CarnegieMellonUniversity-AdClassifier

6 CONCLUSIONS

Our work has extended the available tool-set for domain specific search applications. While there is still an appreciable gap between fully generalizable domain-specific search and the current state-of-art, our work has served to close that gap to a substantial degree by providing insight into scalability of entity resolution, demonstrating utility of general anomaly detection capabilities, identifying sources of model bias, and including multi-modal data in the process, as well as developing tools for extending deep models for image analysis to novel applications, and boosting performance of facial analytics.

7 PUBLICATIONS ORIGINATING FROM THIS WORK

[1] Barnes M, Dubrawski A. “The Binomial Block Bootstrap Estimator for Evaluating Loss on Dependent Clusters”, Conference on Uncertainty in Artificial Intelligence UAI 2017, Sydney, Australia, August 2017.

[2] Barnes M, Dubrawski A. “Clustering on the Edge: Learning Structure in Graphs.” arXiv:1605.01779 [stat.ML], 2016.

[3] Barnes M, Miller K, Dubrawski A. "Performance Bounds for Pairwise Entity Resolution.", arXiv:1509.03302, 2015.

[4] Boecking B, Miller K, Kennedy E, Dubrawski A. "Quantifying the Relationship between Large Public Events and Escort Advertising Behavior." Journal of Human Trafficking, 4(3):1-18, 2018.

[5] Qicong C, De Arteaga M, Herlands W. "Canonical Autocorrelation Analysis and Graphical Modeling for Human Trafficking Characterization", 2017. Project report available online at: https://pdfs.semanticscholar.org/2c7e/0604e738a6f3e9c0474e3f5a3bbf283c2d5f.pdf?_ga=2.237013120.100971400.1530572740-1817910455.1530572740.

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

[7] Girdhar R, Fouhey D, Kitani M, Gupta A, Hebert M. "Cutting through the clutter: Task-relevant features for image matching." In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, pp. 1-9. IEEE, 2016.

[8] Nagpal C, Miller K, Boecking B, Dubrawski A. "An Entity Resolution Approach to Isolate Instances of Human Trafficking Online", Workshop on the Noisy User-Generated Text at the EMNLP 2017: Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, September 2017. (preliminary version published on arXiv:1509.06659 [cs.SI], 29 Jan 2016.)

[9] Rabbany R, Eswaran D, Dubrawski A, Faloutsos C. "Beyond Assortativity: Proclivity Index for Attributed Networks (ProNe)", Advances in Knowledge Discovery and Data Mining, Kim J, Shim K, Cao L, Lee J-G, Lin X, Moon Y-S (Editors), Lecture Notes in Artificial Intelligence 10325:225–237, Springer 2017. (presented at the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2017)).

[10] Rabbany R, Bayani D, Dubrawski A. "RedThread: Active Search of Connections for Case Building and Combating Human Trafficking Online", Knowledge Discovery in Databases (KDD) 2018, London, UK, August 2018.

[11] Rabbany R, Bayani D, Dubrawski A. "Active Link Inference for Case Building Investigation". NIPS Workshop on Advances in Modeling and Learning Interactions from Complex Data, Long Beach, CA, Dec 2017.

[12] Wang X, Shi Y, Kitani K. "Deep supervised hashing with triplet labels." Asian Conference on Computer Vision ACCV 2016.

[13] Wei S-E, Ramakrishna V, Kanade T, Sheikh Y. "Convolutional pose machines." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR 2016.

LIST OF ACRONYMS

CHT Counter-Human-Trafficking

ATF Bureau of Alcohol, Tobacco, Firearms and Explosives

CAA Canonical Autocorrelation Analysis

CCA Canonical Correlation Analysis

sLDA Supervised Latent Dirichlet Allocation

GATE General Architecture for Text Engineering

ROC Receiver Operating Characteristic

SUG Surrogate User Group

CP# Challenge Problem #