Data Integration from Open Internet Sources and Network Detection to Combat Underage Sex Trafficking

Eduard Hovy
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA 15213, USA
hovy@cmu.edu

Nicole Marie Bryan
Montclair State University
Montclair, NJ 07043, USA
bryann@mail.montclair.edu

Andrew Philpot
University of Southern California
Information Sciences Institute
Marina del Rey, CA 90292, USA
philpot@isi.edu

Daniel Ribeiro Silva
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA 15213, USA
drsilva@cs.cmu.edu

Abhishek Sundararajan
Metaio GmbH
abhishek8931@gmail.com

ABSTRACT

Victims of sex trafficking are compelled into sexual exploitation by the use of coercion, deception or fraud, and often find themselves forced into a situation of imprisonment and slavery. This form of human trafficking is a serious felony and law enforcement agents have active ongoing work to combat such crimes, with especial focus on under-age sexual trafficking. Sexual exploitation from traffickers can be often masked as services such as classified ads and escort and massage services (EMS) on the Internet. In this paper, we describe the prototype of a law enforcement support system that aims to make daily extractions of online data, filter relevant information, find hidden patterns and display relevant leads to law enforcement agents. The system uses information retrieval and integration, natural language processing, image analysis, and data linking techniques to allow various forms of relevant information visualization for supporting the combat of sex trafficking. It has been used by law enforcement agencies on specific occasions, and is being developed to suit certain operational needs.

Categories and Subject Descriptors

H.3.1 [**Information Stogare and Retrieval**]: Content analysis and Indexing; 1.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Design, Theory, Measurement

Keywords

Sex trafficking, Child trafficking, Prostitution networks, Law enforcement, Information Integration, Natural Language Processing

1. INTRODUCTION

The illicit use of people against their will is a pernicious problem, especially so when the exploited are children. Every year in the US, numerous children run away from home, only to fid themselves destitute and desperate after a few weeks or months. As described in [7, 8], these children are frequently picked up by a seemingly sympathetic older person, who systematically 'grooms' them and eventually forces them into prostitution, often against their will, and in such conditions that they are unable to escape. Law enforcement agencies around the country, at Federal, State, and local levels, are increasingly devoting people to combating this problem [5]. Task Forces in Los Angeles, New York, and Chicago regularly draw together place, DAs, social workers, psychologists, and NGOs to assist with the problem.

While most forms of slavery and trafficking are covert, sex trafficking includes one public point: where the services are advertised. This is where law enforcement officials attack the problem, using sting operations to try to locate and save underage purveyors of illegal sex. Unfortunately, a large huber of advertisements are published on the internet and in periodicals every day. And since law enforcement units are underfunded and often required to take time to address other pressing problems, they struggle to handle the overwhelming flood of such advertisement efficiently, in order to locate the underage and likely most-needy of help.

Our technology is designed to help alleviate this problem. Our systems automatically harvest relevant web content, extract and systematize and store the relevant information, make estimates about the age of the poster using natural language and image analysis techniques, and retain only those postings that are deemed likely to be from underage posters — less than 2% of the daily intake. For purposes of enforcement action, we focus at this point on outcall (client hotel based) modalities. This information is the arranged in various ways and made available to law enforcement officers to facilitate their search and rescue operations.

In this paper, we first describe the information harvesting and then the information massaging and display procedures. In Section 3 we briefly describe some interactions with law enforcement officials.

2. APPROACH

Our system is divided into two distinct but consecutive blocks. The first one (WAT - Web Archival Tool) the successor to TrafficBot [11] crawls websites containing escort/massage service ads, to automatically identify people who might be underage children. The second block uses gathered data to build prostitution graphs and track group migrations across the country. Data crawling and extraction are organized around three-level geographic metadata (region, metro area, city/neighborhood) to better align different city-based crawl results and to better track movements among and between metro areas. We developed a filtering mechanism to block out all material from people not of interest to law enforcement, since we focus on children only. Increasingly, we are focusing on multidimensional data analytics, combining time, location, phone usage, and other data to identify trends that might help uncover groups of children being 'run' by the same trafficker.

2.1 Data Gathering and Canonicalization

To obtain information indicating possible cases of human trafficking, WAT performs a daily targeted web crawl focusing on sites which include advertisements for escort services. At this writing, WAT extracts data from six seed sites: Backpage, Cityvibe, Eros, Humaniplex, MyRedbook, and SugarDaddyForMe.

For each site, a specific crawler has been developed following a general protocol:

1. Identify all location-specific entry portals of interest. For example, Backpage features about 500 subdomains such as lasvegas.backpage.com; while Cityvibe uses a directory structure to localize its offerings, such as www.cityvibe.com/lasvegas.
2. Either according to a predetermined schedule or on-demand, the crawler will be launched on each such portal on a given host. For our standard areas of interest, we crawl them three times per day each from two sites, rotating the host every day, using a bank of four local and four cloud-based hosts. Spreading the crawling around different IP addresses at difrerent times provides better coverage in the cases of throttling by sources and/or local host failures.
3. For each such entry site, retrieve asset of ad-level URLs of interest. In some cases, the entry portal is essentially an index list. In other cases, the crawler may need to following "next page" links and/or filter links for content of location.
4. Extract the pages themselves using GNU wget, retrieving all related assets such as images and stylesheets, pruning links to sibling pages, and then rewriting all site-derived links to resolve to local copies. For example, if a Backpage-hosted ad refers to images hosted on backpage.com as well as on adultfinder.com, the backpage image data will be cached locally and relinked to the local caches, but the adultfinder.com links will be left as is. This is to establish a reasonable snapshot for the relevant site contents for later inspection.
5. After one day's crawls are complete, the raw wget results are archived to long-term storage and are also combined into a single daily bucket folder. In this step, repeated visits to the same page on the same day are represented by the artifacts of the last visit.
6. Given a daily crawl folder, a site-aware extraction script opens the file corresponding to each individual advertisement. Using a combination of robust HTML parsing (using BeautifulSoup [10]) and regular expressions, a subgraph of objects is constructed corresponding to the content in the post. Figure 1 contains a simplified entity/relationship diagram correspond to the data model in use. In particular, the text of the advertisement is broken into three Text objects, one each for the ad copy (body), the title, and the field indicating location of the advertiser. Text is tokenized using a NLTK text tokenizer [2] customized to treat the typographical and content characteristics of the domain, including treating all Unicode characters above the Basic Multilingual Plane as individual tokens. Phone numbers, often obscured in the ad copy using spelled-out numbers, I/O for 1/0, and other tricks, are canonicalized using a chain of regular expression rewrite rules. Tokens are compared against a set of known working names, including both typical given names and those characteristic of the domain. Because images constitute about 90% of the content by disk size, when an image is encountered whose file contents hash signature has been seen before, the file contents are replaced on-disk by a symbolic link to save space. We model the object subgraph using SQLObject [4], back-ending into MySQL for permanent storage.
7. Once the surface-level extraction is done, first-order analytics are executed. Chief among these is determining ads related to other ads. Crosslinks ("other ads by this poster") included on some ads on some times are extended with ads found using the following mechanisms:
   (a) Each ad body text is inserted in an Apache Solr index [1] and the Solr MoreLikeThis query us used to compute near neighbors. Neighbors within a thresholded distance of a given post are deemed related.
   (b) Ads sharing at least one facial image, or a high proportion of all images, are deemed to be related.
   (c) Ads sharing phone numbers are related. Additional analytics on extracted objects are performed at this time as well. Using OpenCV [3], images are marked as banner text images, images of persons, or other. Images depicting more than one person are annotated as such (this may indicate two persons trafficked together). Post body texts

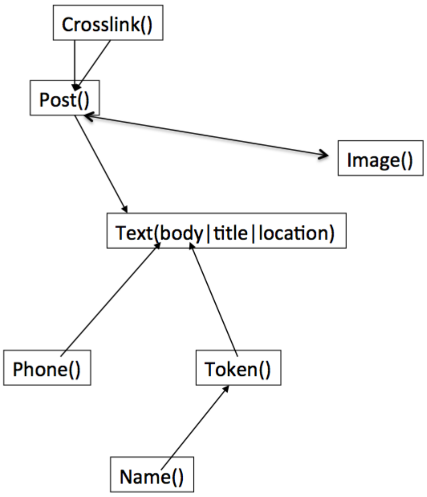


Figure 1: A simplified diagram for the entity/relationship used in the data model.

are analyzed using machine learning classifiers [9, 6] as well as ad hoc keyword fragment pattern matchers, to detect situations of particular interest, such as services offered (massage, spa, outcall, etc.), provider requests (age or racial preference of client), and any such determinations are added to the stored data model.

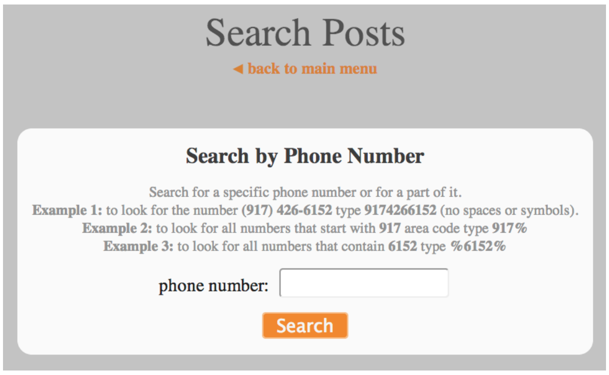
8. Filtering of posts and associated images and other records to remove out-of-domain documents may be done either after step 6, using a simple surface filter, or using the more comprehensive annotations available after step 7.

2.2 Analytics and Visualization

The main purpose of the analytics tool is to consider the entire data history in order to find underlying patterns and existing relations between different postings and posters, or to find typical patterns of activity of posters. In order to do so, we have developed a system that constantly analyzes the collected data and is able to detect relations between postings and/or posters. The links discovered are tracked by the system and continuously inserted into one of several graphs. Each graph represents a distinct active prostitution network (also referred to as cluster or group below).

In order to identify the individual groups, we use a graph clustering approach that is very strict about the criteria for linking posts and posters. As a consequence, any detected link on the data will almost certainly represent a real connection in real life. The system also includes a cluster merge feature. Every time we obtain a new information node that is linked to two different existing networks, the networks are merged and considered to be a new, wider network. A record is kept of how the merge occurred. All cluster and merging information is processed offline and stored in a database that allows rapid construction of any existing cluster.

Such a structured storage of clusters offers an opportunity for many possible analyses. We have focused on analyzing migration patterns of different clusters across the country. Since the contents of each cluster includes the location and the creation timestamp of each posting, we are able to map and keep track on real-time how each cluster is moving. The migration computation is also processed offline and stored in a structure that allows for fast mapping of any cluster migration.


Search Posts

Search by Phone Number

Search for a specific phone number or for a part of it.
Example 1: to look for the number **(917) 426-6152** type **9174266152** (no spaces or symbols).
Example 2: to look for all numbers that start with **917** area code type **917%**
Example 3: to look for all numbers that contain **6152** type **%6152%**

phone number:

Figure 2: Part of interface for querying database

The system contains three integrated interfaces, as described below.

2.3 Interface for Querying Database Postings

A simple query interface has been implemented to search the database for recorded postings. Figure 2 shows the part of that interface that allows users to make queries related with phone numbers. The search tool includes:

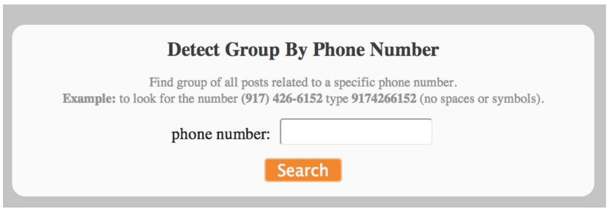
- Search by phone number (entire number, number fragment or region code).
- Search by city
- Search by date
- Combined search (multiple criteria)

The main utility of this tool is to perform background checks on information that is known a priori to be potentially useful.

2.4 Interface for Cluster/Group Detection and Analysis

This interface was implemented to visualize the prostitution clusters that are detected and recorded by our system. It contains two main features:

- Seeded cluster detection: This tool is particularly interesting if one wished to dig deeper into some existing information or lead. By seeding the system with some information (such as a phone number, a poster, or a specific posting), the user can locate the cluster that contains that information and shows all information associated with that specific prostitution network (e.g., related phone numbers, posters, postings, timestamps). Figure 3 shows part of that interface and Figures 4 and 5 show how part of the results are displayed.



Detect Group By Phone Number

Find group of all posts related to a specific phone number.
Example: to look for the number **(917) 426-6152** type **9174266152** (no spaces or symbols).

phone number:

Figure 3: Part of interface for detecting phone-seeded cluster

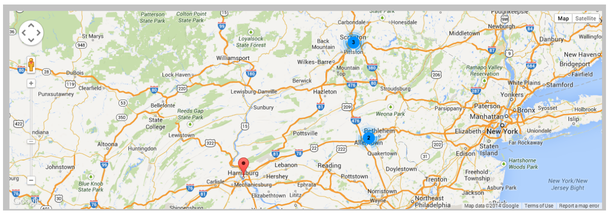


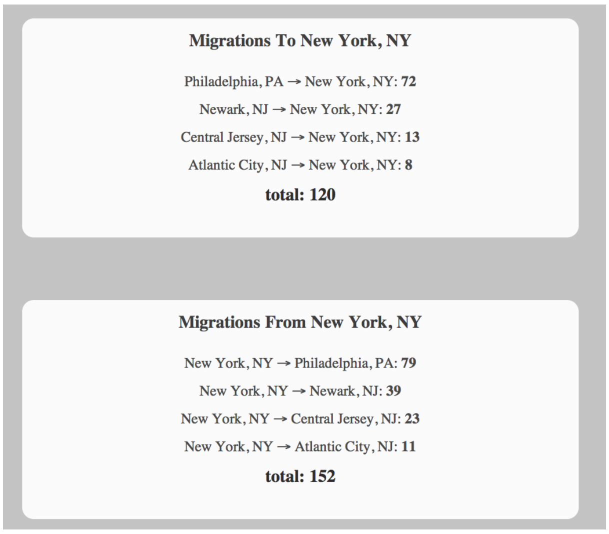
Figure 4: Integration with Google Maps API for easier visualisation of clusters

- Clusters Overview: In case there is no particular seeding information available, the system also contains a tool for simply going through the existing clusters. The system will display a list of the networks detected so far with some key information about each specific cluster. The information includes:
  - Size of cluster: total number of postings comprised in the network.
  - Number of posters: total number of posters in the network.
  - List of names: the system contains a name detection feature that automatically detects and extracts names from the posting text and displays it on the system.
  - Name-based age estimation: We have incorporated into the system government-available data about the popularity of names for newborn girls in the USA across the last years. We combine the name-extraction tool with this information to estimate the girl's names.
  - Cluster geography: list of cities where the cluster was detected.

2.5 Interface for Group/Cluster Migration

This interface was implemented to visualize the migration movements of the detected prostitution networks around the country. The visualization is based on a timeframe selected by the user. The interface displays all cluster migrations that occurred during the selected time window. This is particularly interesting if we wish to analyze the dynamics of migrations during a specific event or period in the past or if we want to keep a daily tracking and monitoring of cluster movements.

The interface also allows the selection of a list of cities of interest for the migration analysis. This implies that the system will only display migrations that occurred from or to one of the selected cities. The interface is divided into two blocks:



Migrations To New York, NY

Philadelphia, PA → New York, NY: **72**

Newark, NJ → New York, NY: **27**

Central Jersey, NJ → New York, NY: **13**

Atlantic City, NJ → New York, NY: **8**

total: 120

Migrations From New York, NY

New York, NY → Philadelphia, PA: **79**

New York, NY → Newark, NJ: **39**

New York, NY → Central Jersey, NJ: **23**

New York, NY → Atlantic City, NJ: **11**

total: 152

Figure 6: Part of the results for a selected cluster.

- Migration statistics: This first block of the interface displays a summary of all the migrations that occurred for the selected cities of interest. As an example, if the user selects New York as the city of interest, this part of the interface will display the migration numbers that occurred, in the selected timeframe, from each tracked city to New York, as well as the numbers from New York to each tracked city (as long as that number is positive).
- Cluster details: The second block of the interface lists all the clusters that participated on the migrations for the selected timeframe and cities of interest. As in Interface 2.4, this interface displays a list of each one of the clusters, with some of their key information.

Figure 6 shows the first block of this interface, with the statistics for New York City.

3. APPLICATION

Nicole's text here

4. ACKNOWLEDGMENTS

XXXXXXXX

5. REFERENCES

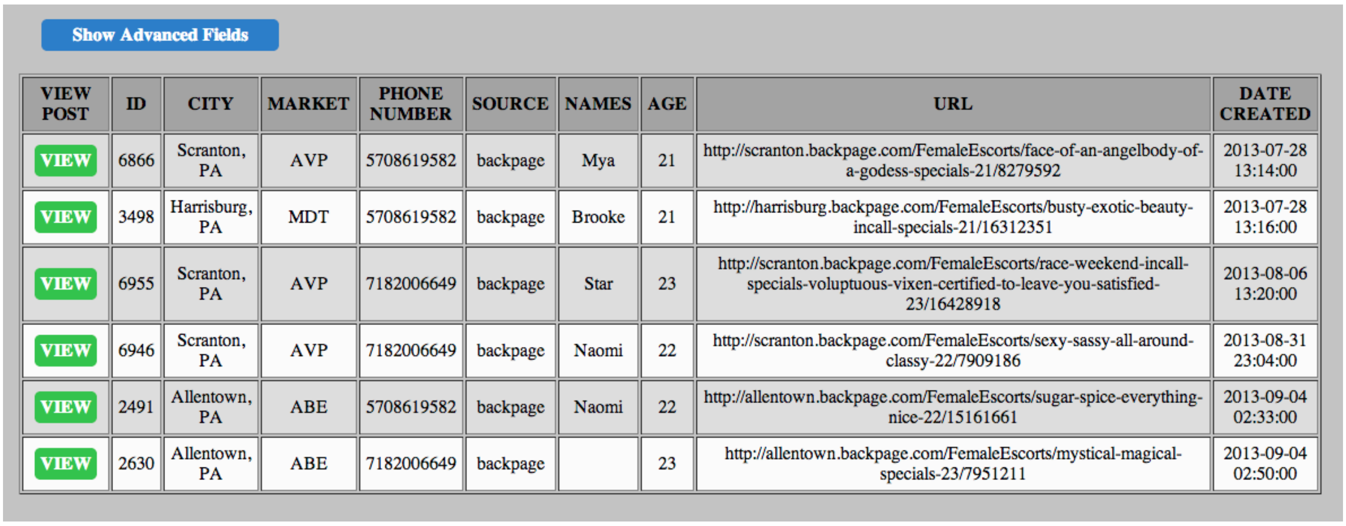
[1] Apache solr - http://lucene.apache.org/solr/index.html, Feb. 2014.

[2] Nltk 3.0 - http://www.nltk.org/api/nltk.html, Feb. 2014.

[3] Opencv - http://opencv.org/, Feb. 2014.

[4] Sqlobject - http://sqlobject.org/, Feb. 2014.

[5] D. Banks and T. Kyckelhahn. *Characteristics of suspected human trafficking incidents, 2008-2010*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2011.



VIEW POST	ID	CITY	MARKET	PHONE NUMBER	SOURCE	NAMES	AGE	URL	DATE CREATED
VIEW	6866	Scranton, PA	AVP	5708619582	backpage	Mya	21	http://scranton.backpage.com/FemaleEscorts/face-of-an-angelbody-of-a-godess-specials-21/8279592	2013-07-28 13:14:00
VIEW	3498	Harrisburg, PA	MDT	5708619582	backpage	Brooke	21	http://harrisburg.backpage.com/FemaleEscorts/busty-exotic-beauty-incall-specials-21/16312351	2013-07-28 13:16:00
VIEW	6955	Scranton, PA	AVP	7182006649	backpage	Star	23	http://scranton.backpage.com/FemaleEscorts/race-weekend-incall-specials-voluptuous-vixen-certified-to-leave-you-satisfied-23/16428918	2013-08-06 13:20:00
VIEW	6946	Scranton, PA	AVP	7182006649	backpage	Naomi	22	http://scranton.backpage.com/FemaleEscorts/sexy-sassy-all-around-classy-22/7909186	2013-08-31 23:04:00
VIEW	2491	Allentown, PA	ABE	5708619582	backpage	Naomi	22	http://allentown.backpage.com/FemaleEscorts/sugar-spice-everything-nice-22/15161661	2013-09-04 02:33:00
VIEW	2630	Allentown, PA	ABE	7182006649	backpage		23	http://allentown.backpage.com/FemaleEscorts/mystical-magical-specials-23/7951211	2013-09-04 02:50:00

Figure 5: Part of the results for a selected cluster.

[6] T. Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[7] M. Latonero. Human trafficking online: The role of social networking sites and online classifieds. *Available at SSRN 2045851*, 2011.

[8] M. Latonero and I. Shklovski. *" Respectfully Yours in Safety and Service": Emergency Management & Social Media Evangelism*. International Community on Information Systems for Crisis Response and Management, 2010.

[9] A. K. McCallum. Mallet: A machine learning for language toolkit. 2002.

[10] L. Richardson. Beautifulsoup - http://www.crummy.com/software/beautifulsoup/, Feb. 2014.

[11] H. Wang, C. Cai, A. Philpot, M. Latonero, E. H. Hovy, and D. Metzler. Data integration from open internet sources to combat sex trafficking of minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, dg.o '12, pages 246–252, New York, NY, USA, 2012. ACM.