

**Carnegie Mellon University**

**Predictive Patterns of Sex Trafficking Online**

By

**Emily Kennedy**

H&SS Senior Honors Thesis

Pittsburgh, Pennsylvania

April 2012

## Acknowledgements

Jessica Dickinson Goodman and Matthew Holmes have inspired and supported this thesis from its inception. Jessica started the fire for this project by suggesting using keywords to narrow down and track sex ads for trafficked individuals on the internet across time and space. Matthew provided technical expertise, building the initial version of the program that allowed me to capture and analyze classified advertisements on a large scale. Without the program, this project would never have gotten off the ground.

Dr. Jay Aronson, Director of the Center for Human Rights Science and Associate Professor of Science, Technology, and Society at CMU's Department of History, helped me transition my topic through various evolutions and provided encouragement and advice at key moments of this project. He not only helped me develop my ideas, but also polish my final product. In addition, I appreciate him referring me to Dr. Artur Dubrawski.

Dr. Artur Dubrawski is the Director of the Auton Lab in CMU's School of Computer Science, a Systems Scientist at the Robotics Institute, and Adjunct Professor at CMU's Heinz School of Public Policy, teaching data mining and business intelligence. I cannot express my gratitude enough to Dr. Dubrawski for his willingness to listen to the crazy ideas of an undergraduate and giving me the chance to pursue these ideas. I want to thank him for his direction through the research and writing process and for generously offering the resources at the Auton Lab for use in my research. Without his generosity, my research would not be possible.

Mrs. Saswati Ray was an immense help throughout this project. She worked to automate the data collection process at the Auton Lab, wrote code to clean the data, and wrote code to extract information from the data. She also did preliminary data analysis, taught me how to use TCWI, and helped me when I encountered problems along the way. I am thankful for her valuable work on this project.

I offer my thanks to law enforcement experts who wish to remain anonymous, for their guidance throughout this project. In addition, I would like to thank Dr. Tim Haggerty, Director of CMU's Humanities Scholars Program and Adjunct Professor in the Department of History for his constant inspiration, prodding, and direction. I would also like to thank the HSP seminar team who read through various versions of my thesis and gave me helpful input.

This project was funded by Carnegie Mellon's Undergraduate Research Office and I would like to thank them for their generous contribution to enable my research. These results represent the views of the author and not those of Carnegie Mellon University.

# Predictive Patterns of Sex Trafficking Online

Emily Kennedy

## Abstract

In the past 10-15 years, the internet has become a popular tool for sex traffickers to advertise and sell their victims. For instance, there are thousands of posts each month selling sex on the common classifieds website Backpage.com, some of which may be cases of sex trafficking. I extracted this publicly available data from the U.S. cities represented on Backpage.com. I then used software developed at Carnegie Mellon's Auton Lab to find out if it is possible to detect patterns emerging from the data available in sex ads, such as patterns of travel that traffickers may use. To help identify which posts are more likely to be cases of trafficking, I relied on guidance from law enforcement experts. The research in this paper shows that it is in theory possible to track the movement of similar posts—and therefore, similar pimps or victims—across the country over time. With further development and refinement, the techniques demonstrated in this thesis could become the foundation for a valuable tool for law enforcement to use to prosecute traffickers and rescue victims.

## Introduction

As the internet has become a popular tool, both sex workers and sex traffickers have found the internet useful for advertising escort services. Although the move to the internet has presented new problems for law enforcement in combating sex trafficking, it has also presented law enforcement agents with a useful weapon for fighting this injustice. There is rich data to be gleaned from online ads for sex. This paper will show how a study and analysis of the publicly available data in ads for sex online can be used to gain intelligence on the patterns exhibited by sex traffickers. It can provide supplemental insight to the intelligence that police and federal agencies already have to aid a more accurate and comprehensive study of this problem.

After giving a historical background of the problem of sex trafficking, this paper will discuss why an analysis of this data is preferable to alternative ways of handling the problem. Then, the paper detail how this data was collected, the methodology behind the research, and will

present research results. The research will aim to identify three groups, which may or may not overlap: posts which advertise underage escorts, posts which indicate “shared management” situations (i.e., trafficking), and posts which move across geographic locations. This paper argues that extraction and analysis of this sex advertisement data can help understand patterns of sex trafficking via publicly available online classifieds websites.

### **Definition of Sex Trafficking**

The United States’ Trafficking Victim Protection Act (TVPA) of 2000 defines a “severe form of trafficking in persons” as a circumstance “in which a commercial sex act is induced by force, fraud, or coercion, or in which the person induced to perform such act has not attained 18 years of age” (TVPA 2000). The TVPA was the first U.S. legislation to define an *underage* prostitute as a victim, rather than a criminal. Because the United States defined trafficking victims this way, it made any act of prostitution a minor commits a case of exploitation, rather than prostitution. For those adults who are trafficked for sex, it must be shown that there was one of three factors—force, fraud, or coercion—present in the situation. The important thing to note here is that it is not necessary to show that there was physical restraint against an adult victim in order to term a situation severe trafficking. It is sufficient to show that there was psychological persuasion or coercion towards the individual in order to term a situation severe trafficking. As the Polaris Project—a leading anti-trafficking non-profit—clearly states:

Psychological means of control, such as threats, fraud, or abuse of the legal process, are sufficient elements of the crime. Unlike the previous federal involuntary servitude statutes (U.S.C. 1584), the new federal crimes created by the Trafficking Victims Protection Act (TVPA) of 2000 were intended to address “subtler” forms of coercion and to broaden previous standards that only considered bodily harm. It is important for definitions of human trafficking in the U.S. and around the world to include a wide spectrum of forms of coercion in order for the definition to encompass all the ways that traffickers control victims (“Myths and Misconceptions”).

There are obviously many individuals who engage in prostitution of their own volition. The TVPA does not address those who engage in prostitution willingly, nor is this paper geared towards this group, but rather towards those who are being prostituted against their will, both underage and adult.

## **Background**

### **Vulnerability of Minors**

The TVPA clearly distinguishes underage prostitution as sexual exploitation rather than illegal sex. There are many factors that influence a minor to willingly choose or be coerced into prostitution. In her essay “Beyond a Snapshot: Preventing Human Trafficking in the Global Economy,” Janie Chuang writes, “trafficking has its root causes in poverty, unemployment, discrimination, and violence against women” (2006). Although most young women who are trafficked come from poor or broken homes (TVPA 2000), some are kidnapped against their will, forced into prostitution, and kept there, sometimes by forced drug addictions. This is true in the case of a fourteen year-old girl in Los Angeles who pimp Leo Braggs kidnapped from a city bus, force-fed ecstasy and GHB, and prostituted against her will (Kloer 2010; Goodman 2011). Most young women who are trafficked come from damaged homes and are lured in by the money, shelter, food, and feigned security that the pimps offer; to them, it seems like the best option (Schapiro, Burke, and Sclafani 2005). For the pimps, trafficking young women for sex can be highly profitable. According to a calculation based on one pimp who trafficked four young women and forced them to meet his \$500/night quotas, The Polaris Project—a Washington, D.C.-based non-profit organization—estimated that this pimp brought in about \$632,000 in one year; this is especially lucrative since pimps don’t generally pay income taxes (“Domestic Sex Trafficking”).

## **The Move to the Internet**

Before the internet's popularity, many pimps advertised their victims on the streets. This provided law enforcement with a physical place to attack the problem and find victims. They could do so by doing stings on johns—those who purchase sex—with police officers posing as prostitutes on the street (Desplaces 1992), or fighting trafficking from the supply side, by targeting the pimps themselves. The internet has made it easier and less dangerous to both buy and sell sex trafficking victims. Lauren Hersh, prosecutor and leader of the anti-sex trafficking unit in Brooklyn, NY, says, “Pimps are turning to the Internet. They’re not putting the girls on the street so much” (Kristof, “How Pimps Use the Web to Sell Girls” 2012). Johns don’t have to go out to possibly dangerous prostitute hotspots, and pimps don’t have to reveal their locations as readily as they might when trying to sell their victims in the flesh. The internet also allows pimps to expand their customer base and it’s now common for pimps to move their victims across state lines, further frustrating efforts to rescue the victims (Goodman 2011). In short, the internet has become a tool for pimps to expand their reach for advertising and selling, while maintaining anonymity and security of their geographic location.

## **Craigslist**

Craigslist was founded by Craig Newmark in 1995 as a way for his local community to stay connected for events (“About: Craig Newmark”). Craigslist was one of the first well-known and popular online classifieds websites, and became incorporated as a for-profit corporation in 1999. Although it was expanded to become a medium for buying and selling, Craigslist still prides itself for its “relatively non-commercial nature” (“About: Factsheet”). This is a surprising self-characterization, however, because when Craigslist opened its “erotic services” section, Craigslist creators found a lucrative niche within their own business. Craigslist’s “erotic

services” section, later renamed Adult Services was ultimately shut down after accusations of being a medium for child sex trafficking. It was one of the first sites of its kind to be associated with sex trafficking. It is important to note that not all ads on Craigslist were for trafficked individuals; many of the ads were posted by adult individuals advertising their services of their own volition. Craigslist was projected to bring in \$36 million dollars from these sex ads in just one year, by charging users to post (Stone 2010).

The shutdown of Craigslist’s Adult Services section is a vivid example of the fluidity of this form of advertising. When the Adult Services section was shut down, according to AIM Group, prostitution ads online decreased by more than 50 percent (Kristof, “Where Pimps Peddle Their Goods” 2012). But, in the month after the Adult Services close, the traffic at Backpage.com—a similar online classifieds website—increased by half a million visitors (Stoeffel 2011; Quantcast). According to Quantcast, a web traffic analysis company, Backpage then became the number one website in the United States for facilitating escort ads. This represents the ability of this data to easily flow from one place to another; when one site is shut down, many sellers simply move to other venues.

### **Censorship of Online Classifieds**

Although Craigslist was ultimately shutdown voluntarily, one might wonder, “Why could Craigslist not be prosecuted for facilitating posts that advertise illegal activity?” Because of Section 230 of the Communications Decency Act of 1996, Craigslist, as provider of the advertising service, cannot be treated as the publisher or speaker of what its users write. The Act states: “No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider” (Communications

Decency Act 1996). These are appropriate protections for online classifieds providers, but they result in a vacuum of accountability for the illegal services advertised on the internet.

### **Backpage.com**

Backpage.com—an enterprise owned by Village Voice Media (Kristof, “How Pimps Use the Web to Sell Girls” 2012)—operates under a goal similar to that of Craigslist. On its “About” page on the website, it answer the question of “Why Backpage?” with the response: “Because other people want what you have, and they might have what you want” (“About Backpage.com”). Operating out of Phoenix, AZ and Dallas, TX, Backpage offers a similar feel to Craigslist of a friendly, community website. Backpage currently benefits from the same lucrative business that Craigslist did: facilitating escort ads. According to a consulting company, AIM Group, Backpage makes more than \$22 million per year from its ads for prostitution (both those ads selling the services of consenting adults and those selling trafficking victims) (“Backpage replaces Craigslist as prostitution-ad leader” 2010), and it provides space for about 70 percent of the total online ads for prostitution in the United States (Kristof, “Where Pimps Peddle Their Goods” 2012). Lauren Hersh, the Brooklyn prosecutor and anti-trafficking unit leader, termed Backpage as “a great vehicle for pimps trying to sell girls” (Kristof, “How Pimps Use the Web to Sell Girls” 2012). Nicholas Kristof, who has recently written a number of explosive articles about Backpage for *The New York Times*, commented on the evolution of Village Voice: “Paradoxically, Village Voice began as an alternative newspaper to speak truth to power. It publishes some superb journalism. So it is sad to see it accept business from pimps in the greediest and most depraved kind of exploitation” (“Where Pimps Peddle Their Goods” 2012).

Backpage responded to the accusations made by Kristof with “Statement to Nicholas Kristof RE Backpage January 24, 2012.” In it, Backpage affirmed its agreement with the National Association of Attorneys General that underage sex trafficking was a bad thing. It also cited Ernie Allen, president and CEO of the National Center for Missing and Exploited Children as saying in September of 2011, “Backpage has been aggressively reviewing their ads and trying to remove those ads that are unlawful and suggest they involve the sale of kids for sex.” It mentioned local and federal agencies as applauding Backpage’s responsiveness and cooperation with the problem. In the statement, Backpage went on to accuse many local and national political leaders as using Backpage as a scapegoat to further their political campaigns. Despite this defense by the company, Kristof continued his articles on Backpage. On April 18, 2012, writing about a girl who was trafficked on Backpage at the age of 12, Kristof said, “Backpage cooperates with police and tries to screen out ads for underage girls, but that didn’t help Brianna.” He pointed out the fact that, despite efforts, this problem is still real; and Backpage, facilitating 70% of U.S. ads for sex (The Aim Group, “Sites set combined record for online prostitution-ad revenue” 2012), enables a major share of this problem (Kristof, “Not Quite a Teen, Yet Sold for Sex” 2012).

Trafficking on Backpage is indeed a problem. As of March 31, 2012, Kristof reported, “arrests were made in 22 states for trafficking of under-age girls who had been marketed on Backpage” (“The Secret Owners Behind a Prostitution Website” 2012). For a list of news-documented cases of trafficking on both Backpage and Craigslist, please see Appendix A. Backpage’s role as the most prominent website in the United States for escort advertisements made it the optimal place to start mining data for the research in this paper.

## **Alternatives to Data Extraction and Analysis**

This paper argues that extraction and analysis of this sex advertisement data can help understand patterns of human trafficking via publicly available online classifieds websites. Some might suggest alternatives to an extraction and analysis of this data, such as having the site self-monitor its posts, having the site encourage its users to self-monitor the posts, and shutting down these sections of the site completely; and in fact, these alternatives have already been tried in the case of Craigslist. A look at this recent history will show how these efforts were either largely unsuccessful, or not successful enough to be relied upon as optimal solutions to the problem of online sex trafficking.

### **Self-Monitoring**

#### *Craigslist*

Craigslist claimed in May of 2009 that its Adult Services posts would be “manually reviewed” by Craigslist workers (Michels and Esposito 2009). Craigslist agreed to screen each Adult Services ad for inappropriate or illegal material and report it to the National Center for Missing and Exploited Children. For a company that received thousands of posts *daily* in the Adult Services section in 2009, this was a bold claim to make. Two years after Craigslist agreed to report posts to the NCMEC, amid the hundreds of thousands of posts that passed through the Adult Services section, fewer than 100 posts had been reported to the NCMEC by Craigslist (“CNN: Craigslist and the sex trade” 2010). Due to public outcry, in September of 2010, Craigslist completely removed its Adult Services section and replaced it with the word “CENSORED.” Then it ultimately removed the “CENSORED” label, deleting the section completely (Matyszczyk 2010; Miller 2010).

*Backpage*

Similar to Craigslist's claim, Backpage has agreed to self-monitor the posts in its Escorts section. According to Ernie Allen, President and CEO of the National Center for Missing and Exploited Children (NCMEC), Backpage has been doing a fine job of reporting ads. He said, "Backpage has reported to us 1,600 ads that they believe are suspicious" (Frolik 2011).

Although this is certainly a step in the right direction, Allen says that traffickers can quickly change their advertising tactics to sidestep these filtering methods (Frolik 2011).

In addition, Backpage encourages its users to self-monitor the site. For example, before a user enters the Escorts section of Backpage, the user must click past a warning in which the user acknowledges that he or she is of 18 years of age or older, and also "agree[s] to report suspected exploitation of minors and/or human trafficking to the appropriate authorities" ("Disclaimer").

This manual filtering mechanism is not enough by itself to stop traffickers from advertising their victims. Although it is important to have humans filter through the data at some point, the sheer volume of the data and frequency of posts make it inefficient and impossible for all possible posts to be filtered, let alone compared with the data set as a whole to find possible correlations among posts on a large scale. The research proposed in this paper would allow for large scale analysis of posts as compared to each other, and as compared to incoming posts. Using algorithms to process this data will allow analysts to find large scale, cross-country patterns that are undetectable if posts are merely filtered by individual, unconnected human analysts across the United States. These posts cannot be merely analyzed individually. This research proposes that the most fruitful analysis will come when a post is not only analyzed individually for trafficking or underage prostitution cues, but also when it is analyzed in

comparison with similar posts or posts that share the same phone number in cities across the entire United States.

### **Complete Shut Down of the Adult Services Section**

Another complication regarding Craigslist's involvement in the problem is the dispersal of ads advertising sex. New York Attorney General Andrew Cuomo advocated *against* the removal of the ads, arguing that the presence of the ads would be useful to catch traffickers using information such as IP addresses and credit card accounts (Michels and Esposito 2009).

Microsoft senior researcher Danah Boyd argued that the Adult Services section should be reopened for similar reasons (Miller 2010). Once Craigslist removed its Adult Services section completely in 2010, ads appeared to merely have dispersed to other sections of the site such as "casual encounters" (Wright 2010) and "therapeutic services," (Matyszczuk 2010) making them even harder to track. In addition, ads seem to have migrated to Backpage.com, judging from the half million increase in visitors that occurred in the month after the shutdown of the Craigslist Adult Services section (Stoeffel 2011; Quantcast).

Although shutting down Backpage's Escorts section would momentarily disperse or stop sex ad postings, it is not a long-term solution to the problem. Given the lucrative nature of the business—as evidenced by the more than \$22 million that Backpage is estimated to bring in annually (Kristof, "Where Pimps Peddle Their Goods" 2012; The Aim Group "Backpage replaces Craigslist as prostitution-ad leader" 2010)—it is not outlandish to expect that, within time, another Backpage-type service would pick up the client base; in fact, it is possible that a future Backpage successor could develop more sophisticated protections against shutdowns. With this information in mind, it is important that action be taken *now* to mine the depths and

discover the intelligence that an analysis of this publicly available could provide to law enforcement, with the aim of prosecuting traffickers and rescuing victims.

### **Prior Work**

One of the first studies done on applications of data mining for online human trafficking was done in 2011 at the University of California's Annenberg Center on Communication Leadership & Policy. In a report published by a team led by Dr. Mark Latonero in September of 2011, "Human Trafficking Online: The Role of Social Networking Sites and Online Classifieds," the team did an analysis of some of the data on Backpage. For example, the report looks at frequency of female escort posts leading up to the Super Bowl XLV in 2011 in Dallas, TX. The team began the research with the hypothesis that female escort post frequency would increase in this city leading up to this event. By gathering and analyzing the data, the team was able to confirm this and they were also able to determine that "the Super Bowl attracted a slightly older pool of reported ages than usual" (Latonero, Berhane, Hernandez, Mohebi, and Movius 26), but they were not able to isolate cases of sex trafficking from among the data.

In the team's analysis of the frequency of Backpage adult section posts in Los Angeles, it was confirmed that although data mining could be useful for narrowing down pools of relevant posts, research would be more fruitful with the participation of a human analyst to provide context not represented in the data, and to ultimately make decisions regarding the available evidence. The team then went on to identify potential areas of technology that could be developed or used to further aid in refining a data analysis of sex ads for trafficking (Latonero et al. 28-30).

## **Human Rights Guidelines**

Latonero et al. outlined a number of human rights guidelines that are recommended and will guide the research in this paper.

- 1) The goal of any innovation in analysis of this data should be for the benefit of sex trafficking survivors and victims.
- 2) Companies that facilitate trafficking as a result of their use—such as Craigslist or Backpage—should be aware of this problem and work to innovate solutions to the problem, as well as cooperate with non-profit, for-profit, and governmental organizations that are working to solve this problem.
- 3) Due to the frequently changing and developing nature of technology, commitment and participation in solving this problem in the technological context must be constant.
- 4) When this data is extracted and analyzed, it must be done with an awareness and concern for the ripple effects on trafficking survivors and victims, as well as individual rights and privacy (Latonero et al. v-vi).

## **My Research**

In order to determine whether it is possible to find meaningful patterns of trafficking in Backpage sex ads and to locate trends that may be useful for law enforcement agents, this research aimed to identify three types of ads, which may or may not overlap: posts which advertise underage escorts, posts which indicate “shared management” situations (i.e., trafficking), and posts which move across geographic locations.

### *Assumptions*

The importance of this analysis is supported by some foundational assumptions:

- 1) Sex trafficking is viewed as a supply and demand mechanism. This may be useful for identifying cases of underage trafficking.
  - a. Because youth is a selling point for many buyers, in order to maximize impact and profit from an ad, the seller must communicate this selling point to the buyer through the ad. Since it is possible for a potential buyer to detect this selling point, it should be possible for a researcher or law enforcement agent to also detect this selling point in many cases.
- 2) Analyzing the text of this data may be useful for isolating cases of trafficking—of either underage or adult victims. There are certain textual cues—explained in detail later in the paper—that may tip off an investigator to a case of sex trafficking.
- 3) Analysis of the data is most useful if it is used in conjunction with law enforcement's prior intelligence about an illegal activity. An analysis of the data in itself cannot say with one-hundred percent certainty that a certain post is an underage trafficking situation. But, if the selection of posts could be narrowed down to those that are more likely to be underage or trafficking situations or both, then law enforcement might be able to use the prior intelligence they *do* have in a more focused and efficient manner.
- 4) Law enforcement will be most effective in detecting traffickers and trafficking victims when it is *anticipating* the tactics, decisions, and movements of the individuals, rather than *reacting* to them. Because there is not enough manpower to read through the thousands of posts that appear in the Escorts section on Backpage.com daily, data mining has the potential to be an extremely useful tool for automating the mundane task of

sifting through posts. This research hopes to use publicly available data to increase law enforcement's capacity to identify and anticipate the strategies and movements of traffickers.

- 5) New technologies used for media transfer will be increasingly used for trafficking as they offer traffickers ease and perceived anonymity of posts. Because of this, law enforcement should seriously consider the implementation of these technologies in order to fight traffickers where they are operating.

### Data Collection

Below is a real post sample from Backpage.com. The title is in bold at the top and usually ends with the poster's reported age. The post also lists a time stamp, the poster's age, location, and poster ID. The description usually lists some attributes of the individual(s) and sometimes a contact phone number. Note the emphasis of the ad on the individual advertised as arriving directly from Vietnam. This could potentially be a sex trafficking indicator. The reported age of the individual is not underage, but is on the younger side of escort ages. In addition, language in the ad emphasizes the newness of the individual. This could be a trafficking indicator, or merely an indication that youth and novelty are common advertising points for escort services.

#### **Early delivery from Vietnam YOUR HOTTEST asian has arrived!!! - 21**

Posted: Friday, April 20, 2012 4:52 PM

---

Your delivery from Vietnam is HERE. Come unwrap YOUR NEW Hot arrival and see just how sweet and fun she is. Promise she is MOST hot and sweet and you WILL be MOST happy and smile LONG TIME

714-678-██████

Poster's age: 21

- Location: Orange County, 22 Fwy Near Brookhurst
- Post ID: ██████████orangecounty

[Email this ad](#)

Many ads on Backpage.com are posted by individuals selling their services without coercion or force. In contrast to the previous ad, the following ad does not exhibit many common trafficking triggers. It seems to be the case of an adult woman selling herself without any apparent intervention. The woman's age is listed as 25 years old, and the ad emphasizes that not only does the individual have "well reviewed" reputation, but the ad also emphasizes that the individual is working independently.

[backpage.com](#) > [new york adult entertainment](#) > [new york escorts](#)

• **VERY BUSTY** • **EXOTIC BEAUTY** • **GORGEOUS CURVES** • - 25

Posted: Wednesday, April 25, 2012 12:03 AM

Reply: [click here](#)

Hello Gentlemen . . .

Are You Looking For A Sexi, Hot, And Busty Girl?? Then Feast Your Eyes On ME!! Im Everything You Dont Want To Miss Out On And One **\*\*Exotic Beauty\*\*** You'll Never Forget!!

•5'3•135•Italian & Latina BrUnEtTe/Black Hair •38DD- all natural •curves in all the right places •Enchanting Eyes  
•Real/Recent Pics

☛ GIVE ME A CALL ☛ 401..458. \$100 specials.. \*NO TEXTS/PRIVATE CALLS

Well Reviewed 100% Independent

Poster's age: 25

• Location: Long Island,

• Post ID: newyork

[Email this ad](#)

### *A Note on Reported Age*

Because the TVPA defines sex trafficking as the prostitution of any individual under the age of 18, it is clear that a pimp trying to write an ad masquerading as an ordinary escort service would not explicitly state that the age of the poster is under 17, even if it is. Law enforcement agents such as Lauren Hersh concede that listed ages on Backpage ads may not always be what they seem. She says of the reported ages of victims on Backpage, "I see 19 and I immediately think 13" (Kristof, "How Pimps Use the Web to Sell Girls" 2012).

## Methodology

Given that there are hundreds of posts in the Escorts section of Backpage across the United States per day, it was completely unfeasible to collect this data manually. Matthew Holmes, a Pittsburgh-based software programmer, and Jessica Dickinson Goodman, an anti-human trafficking activist and social media consultant, generously offered their time and help in implementing a data collection system. They had worked previously to provide an interface for non-profit and law enforcement groups that would narrow down sex ads on Craigslist that were likely to be trafficking by looking for keywords such as “sweet” or “adorable” and counting them in order to give analysts an idea of the impact of these ads (Dickinson Goodman and Holmes 2011). Holmes wrote a program that would receive data from the Backpage RSS, a feed made for users to be able to receive automatic updates from their favorite websites; applications such as Google Reader compile RSS feeds to group data from a number of blogs and websites in one location. The program recorded data from every Escorts page for the 443 pages representing each United States region or city listed on the site. From each post, the program would gather the title of the post, the text within the post, the posting location (the specific Backpage site of the city or region it was posted in), a more specific location if specified, and a time stamp recording exactly when the information was posted, including the day of the week, year, month, day, hour, minute, and second. This information was extracted from all Backpage sites and saved in comma separated value format, a file format that separates post data into categories separated by commas, which allowed it to be easily uploaded into various analytic software tools, including Microsoft Excel. This large scale data collection commenced in late October of 2011 and is ongoing. For the purposes of this thesis, data will only be analyzed from October 2011 to April 17<sup>th</sup> 2012.

*The Auton Lab & the Time-Series Cube Web Interface*

In order to approach this research from a large scale data analysis perspective, Dr. Jay Aronson referred me to Dr. Artur Dubrawski for his expertise. Dr. Dubrawski is the Director of the Carnegie Mellon Auton Lab, a research group within the CMU School of Computer Science. The Auton Lab's goal is to research new methods and applications for statistical data mining. As it turned out, the programmers and scientists at the Auton Lab had already developed an efficient and robust system for analyzing large, messy data sets, among them, the Time-Series Cube Web Interface (TCWI). TCWI is set apart from conventional database tools in that it allows for quicker statistical analyses as well as faster data retrieval and "slicing and dicing" of data. TCWI allows the user to quickly make queries, for instance, narrowing down the data to only contain posts in a certain state or containing certain words or phrases. TCWI is revolutionary for its ability to search data sets containing hundreds of thousands of entries in a matter of seconds. Queries on the data set can be useful to sort posts, search for phrases within posts, find all posts containing a certain phrase, or quickly count posts that fall in a certain category. In addition, it allows the user to model data in various ways on the spot, providing an efficient and communicative mechanism to visualize the data (Sabhnani, Dubrawski, and Moore; Ray, Michalska, Sabhnani, Dubrawski, Baysek, Chen, and Ostlund 2008).

In order to expedite the data collection process, Saswati Ray, a research programmer at the Auton Lab automated the data collection process so that the code would run automatically at regular intervals and be saved on secure servers at the lab. In addition, Ray wrote a program that would automatically clean the data to reduce superfluous information that would hinder an efficient analysis. This code would clean the data by taking out insignificant, yet common words such as "the" and "and," which would slightly reduce the data size and expedite the query and

analysis process. This cleaning process was refined over time, for instance, when we discovered that it would be important to keep instances of both the words “girl” and “girls.” This was deemed to be important because the instance of the word “girls” (versus the singular) could be evidence that would indicate a trafficking—a.k.a. shared management or pimping—situation. In addition, it would correct some common spelling mistakes as well as delete postings that had been duplicated during the data-gathering process, i.e., posts that had the same exact time stamp and information. Note that this process would not delete posts that had the same exact information within them but had been reposted at different times by the author. It is common practice for posters in busier cities to repost their ad multiple times a day in order to stay on the first page of the site.

### *Problems with the Data*

Data of the type that we collected in this project (i.e., large amounts of textual content created by many individuals who are engaging in potentially illegal activities) can be difficult to compare and analyze for several reasons. Some examples of these roadblocks are misspellings of words, multiple spellings of names, and difficulty of parsing phone numbers that are typed in a mix of words and numbers (e.g., “one52fivesix99”). Typically, posters do this to prevent automated programs from extracting phone numbers to put them on sales call lists. Other problems have to do with the accuracy and honesty of the data being posted. It is important to acknowledge that data posted by the author may be intentionally inaccurate, such as the poster’s name, age, physical description, and even photo. Another thing to acknowledge is that, although a poster may advertise their services in a certain city, this does not guarantee that they are physically present in that city at the time of the posting. For instance, a pimp could advertise for the same victim in three different cities across the country in the same evening, with the

knowledge that he is willing to pay for a last-minute plane ticket to fly her out if the profit she would make that night outweighs the plane ticket cost (Interview with Anonymous law enforcement agent 2011).

At this initial point in data analysis, we decided that it would be most beneficial to start by only analyzing the text of the data, rather than also taking photos into account. Adding photos to our data collection would lead to data storage limitations and the need for tailoring preprocessing, and analyzing them would require a potentially laborious expansion of the currently available system. Even though this research does not include an analysis of photographic data, it is an important and promising area that demands further research and inquiry.

### **Data Analysis**

The research in this paper is intended for use in tandem with law enforcement intelligence. It could be used to narrow down the data to posts that are most promising at potential cases of underage prostitution or sex trafficking. The research will aim to identify three groups, which may or may not overlap: posts which advertise underage escorts, posts which indicate “shared management” situations (i.e., trafficking), and posts which move across geographic locations. The following section will lay out the methodology used for narrowing down identification of these groups and events.

#### *Pinpointing Underage Escorts*

Identifying underage escorts is crucial for victim identification, given the Trafficking Victims Protection Act’s definition of a “severe form of trafficking.” In addition, as mentioned before in the “Assumptions” section, youth is a selling point for many buyers of sex. Although we cannot know with 100% certainty that the data in a certain post is completely accurate, that

data can still be useful to narrowing down a data set to contain posts that are more likely to be cases of underage escorts. One method of doing this was to take into account weight and bra cup size of the individuals advertised. Although many of the ads did not contain this information, some did, and Saswati Ray of the Carnegie Mellon Auton Lab wrote code that could extract this data from a post (if it was available) and create another category to store this information for each post. This would essentially add another column into the spreadsheet of data for weight, if it was mentioned; it would add another column for cup size, if mentioned. This information is useful with the assumption that smaller cup sizes and lower body weights would correspond to smaller—and possibly underage—girls. For instance, it would be likely that a girl advertised as being 100 lbs. (as has been seen within our collected data set) was not older than eighteen.

Given that the data collected has a textual form, emphasis was also placed on identifying words or phrases that could indicate an underage escort, while also taking into account the reported age of the poster. A CNN Report on sex trafficking in hotels confirms the connection between a demand for underage girls and the advertising necessary to sell them. Kimberly Ritter, a meeting planner who works with the hotels to fight the sex trafficking that happens on their premises said in the CNN interview, “Escort ads posted online do not obviously state that sex with children is being sold...but customers who want children know to look for words like ‘fresh,’ ‘candy’ and ‘new to the game’” (Hetter 2012). In a similar way, this report set out to narrow down cases of underage trafficking via a textual analysis of the data. Law enforcement experts as well as grassroots anti-trafficking activists with expertise in this area were consulted for identifying which phrases were important to pay attention to.

### *Potential Underage Trafficking Example from the Data*

An example of a post that contains some potential underage trafficking triggers is shown below. This post was isolated for a number of reasons. For one thing, more than one girl is listed in the same ad; this could potentially indicate that they are either working together or have the same management. In addition, the listed age in the title is 21, which does not indicate trafficking, but does indicate that the individuals advertised are on the younger side of escort age. Although the age of the first girl is listed, the age of the second is not. The words “sweet” and “girl” emphasize their youthfulness and “new in town” indicate newness which some have correlated with cases of trafficking. Rachel Lloyd, sex trafficking survivor and founder of Girls Educational and Mentoring Services, a non-profit which provides service to sex trafficked girls and young women, detailed this in “An Open Letter to Jim Buckmaster.” In her letter to Buckmaster, CEO of Craigslist, she recounted the story of an 11-year-old girl who was trafficked on Craigslist with the phrase “NEW IN TOWN,” as she was moved by her pimp up and down the East Coast (Lloyd 2010). It can also be noted that the first girl is described as being very slight, which could indicate a young age. Finally, it can be seen that this ad was posted in the Oakland/East Bay Escorts page of Backpage.

[backpage.com](#) > [oakland/east bay adult entertainment](#) > [oakland/east bay escorts](#)

## **Hot Sweet, Gorgeous Asian girl [REDACTED] & [REDACTED] - New in Town - 21**

Posted: Sunday, April 22, 2012 10:28 AM

Reply: [click here](#)

Hi, I am Beautiful & Sweet Asian girl [REDACTED], 22 years old. 5'3", 36D-24-36, 106 lbs. I have smooth and kissable skin, beautiful face, and so sexy body. I'm sweet and playful, my man! I will Rock Your World & Some.

Sexy [REDACTED] - I am new to the Bay Area and ready to meet you. Please call me for happy time.

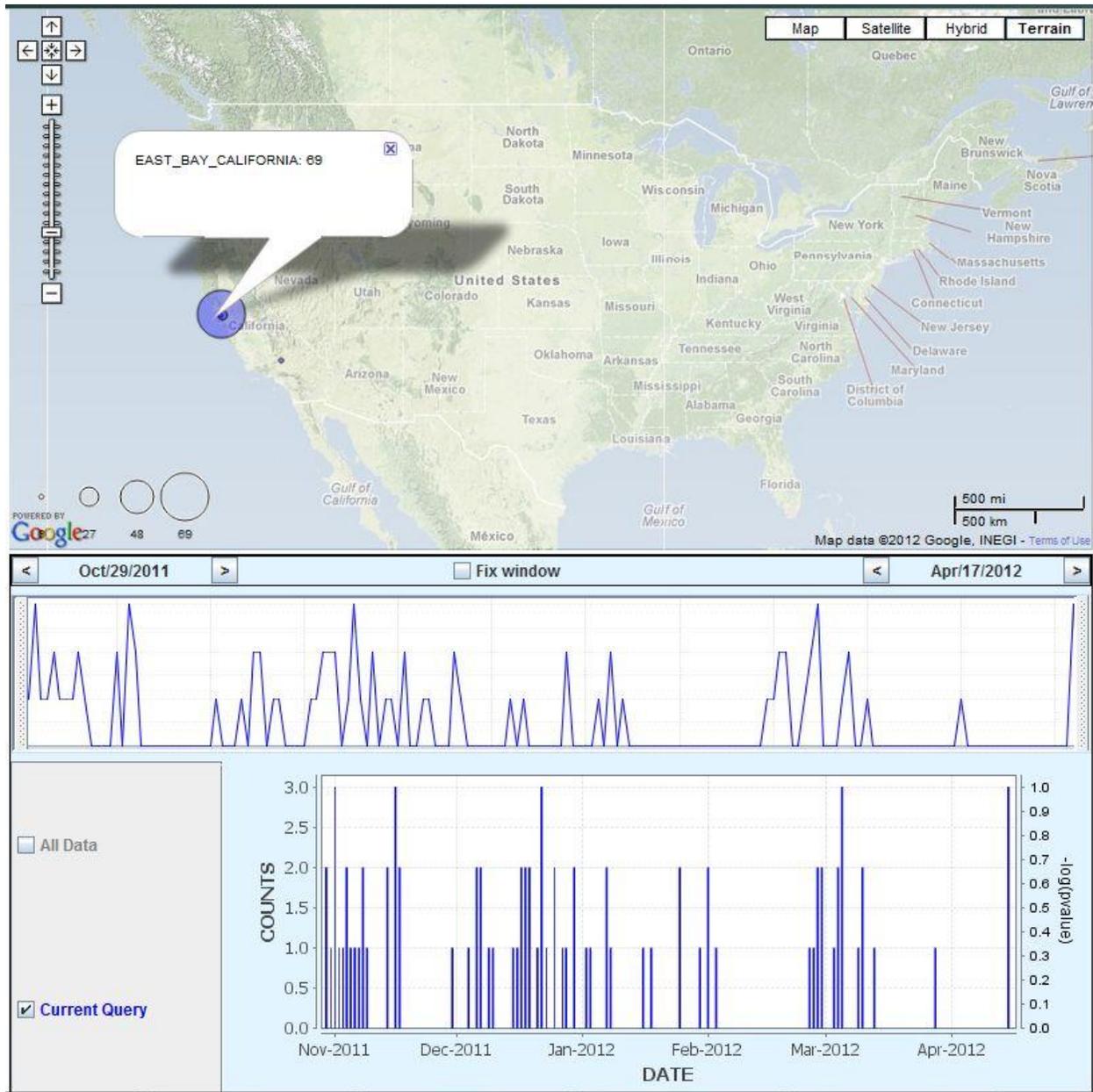
Incall  
160 for 1 hour  
Call for appointment  
510-857-[REDACTED] or 510-586-[REDACTED] no private call.

Poster's age: 21

- Location: East Bay, Union City/Hayward

If an analyst wanted to further investigate this post, TCWI allows the analyst to view the data spatially, across the United States. In an attempt to isolate this post and those similar to it, the following query was used: QUERY=ASIAN=(1), BEAUTIFUL=(1), BODY=(1), FACE=(1), GIRL=(1), NEW=(1), SEXY=(1), SKIN=(1), SWEET=(1), TOWN=(1). That is, TCWI would narrow down the data to only posts that contained all of these words. Note that the posts were not narrowed down by geographic location through the query. Although it might seem that there could be many posts across the United States that contain these words, and that it is not likely that a mere query could narrow down the posts to one geographic location, the query actually did pinpoint the posts to a very specific region, the West Coast, and specifically the East Bay.

After running this query, TCWI represented posts fitting the description in the query on a map of the United States by geographic location. On the map on the next page, one can see circles representing the location of these posts. The larger circles correspond to a greater frequency of posts in that location. The smaller graph below the map displays the frequency of posts in this query over time, from October 2011 to April 2012. Finally, the graph at the very bottom displays more detail about the count of posts over time. One can see that the maximum count of posts in this query in a given period was 3.



It is clear that posts containing these words came from a concentrated region in the United States, around San Francisco, with the majority of posts (69) coming from the East Bay, a subset of San Francisco. There were a smaller number of other posts in this query besides those on the East Bay Backpage: 2 on the California state Backpage, 10 on the Oakland Backpage, and 2 on the San Francisco Backpage. Given the closeness of all of these areas, it is possible that the same group or individual posted these ads on those various pages in an attempt to branch out in

advertising. They appear to have been going on at the time data collection began, and therefore it is likely that the ads were being posted before that time. Although there was a lull around the beginning of April, posts seem to have continued towards the middle of April.

This ability to narrow down data from a textual analysis allows the analyst to pinpoint certain ads to very specific geographic regions, or, in this case, cities. It is important to note that this particular post was isolated to a geographic region without using phone numbers, city names, or names of escorts to identify it. This type of analysis could be especially useful for supplementing other intelligence that law enforcement has, especially when some of the pieces of information are missing.

#### *Pinpointing Trafficked Escorts: Shared Management Situations*

The data was also analyzed in order to narrow down posts that were likely to be cases of shared management, i.e. trafficking. There were a number of ways a post could be identified as potentially a trafficking situation. The following sections discuss some possible indicators of a shared management or trafficking situation, a real example of each, and an explanation of each. Names have been changed and phone numbers edited to protect anonymity.

#### Posts that contain third person language

Example:

*Cece she stand 5'2 tall, 110lbs, 36C-23-35, nice sexy hips and her body is , slim and in perfect model shape with smooth, silky skin*  
("Oakland/East Bay Escorts," Backpage 2012).

This assumes that a post containing third person language is more likely than not to be written by someone other than the escort who is being advertised.

Posts that contain first person plural pronouns such as “we”

Example:

*WE HAVE PLENTY CUSTOMERS but we need GIRLS. Just take some time, come work with us, make what you want, and then stay or leave with PLENTY MONEY !!*

*MAKE 2 STACKS + per week*

*WE OFFER:*

- 1. Your own Phone #*
  - 2. Free advertisement & Photos*
  - 3. Security & screening of your customers*
  - 4. Convenient, Safe, Good neighborhood Location*
- NO FAT GIRLS PLEASE. slim girls, any color/race*  
*(954) 260 XXXX*

*YES, START TODAY.*

*GUYS COME SEE THE NEW GIRLS before they leave ...*  
*(“Miami Escorts,” Backpage 2012).*

This type of language may be used similarly to one describing a business entity and includes words such as “we” and “our.”

Posts that contain the same phone number used to advertise different girls

*347 741 XXXX Jenny Gigi Sonya come on over and experience the hottest girls in town*  
*(“New York City Escorts,” Backpage 2012)*

“Different girls” would be defined as either girls with different names, or different descriptions, or both. This could either indicate trafficking, or it could merely indicate that a number of individuals are working together.

Posts that contain a reference to an outside web site

Example:

*Gorgeous, Prompt, Discreet..*  
*Outcall only! girls pictures 100% real and recent Check our website at adorableplaymates.net No texts please*  
*412-306-XXXX (“Pittsburgh Escorts,” Backpage 2012).*

The presence of a link to an outside website could indicate more elaborate organization for an individual or a group.

Posts that contain similar language shared between two or more ads for different advertised individuals

The key similar phrase between the two examples is in bold.

Example 1:

*HeYY GuYs iM Candy*  
***I am the complete package***  
***An Adorable Princess.***  
***Are you ready to treat yourself***  
***with the company of one sweet,***  
***mysterious delight?***  
*100% Real Pic's & Very Current ~\*~*  
*No Rushing NO GaMes!!*  
★Killer Body !!!  
★Beautiful Smile  
★Awesome Personality  
★A+ Attitude  
CALL ME :) 717-398-XXXX NO BLOCKED CALLS OR TEXT  
THANKS GUYS ;)  
("Seattle Escorts," Backpage 2011).

Example 2:

*HeYY GuYs iM Joline*  
***I am the complete package***  
***An Adorable Princess.***  
***Are you ready to treat yourself***  
***with the company of one sweet,***  
***mysterious delight?***  
*100% Real Pic's & Very Current ~\*~*  
*No Rushing NO GaMes!!*  
*SuGaR FrIeNdLy!!*  
★Killer Body !!!  
★Beautiful Smile  
★Awesome Personality  
★A+ Attitude  
305-773-XXXX ("Miami Escorts," Backpage 2011).

This situation is especially important. This method assumes that if two ads both contain similar language, it indicates that both ads were written by the same person or group of people. If similar language was used to describe different girls—especially different girls in different geographic locations—it could indicate the presence of a common manager or pimp. One thing to note is that it is possible that two ads for two different girls have the same language because, although the two girls are unconnected, one decided to copy the language of the other’s ad. Although this is possible, we assume that it is reasonable to move forward with the original hypothesis and accept a certain small rate of “false positive” discoveries.

Identifying similar language is especially appealing because it could be used to identify similar posts without requiring all of the language in two ads to be exactly the same. Many times, similar language is imbedded in the middle of ads, and therefore hard for the naked eye to identify. This is an instance in which relying on data mining and analysis is very useful in supplementing a human analysis.

#### Posts that advertise for multiple girls in one ad

Example:

*Hi, I am Beautiful & Sweet Asian girl Lulu, 22 years old.  
5'3", 36D-24-36, 106 lbs. I have smooth and kissable skin, beautiful  
face, and so sexy body. I'm sweet and playful, my man! I will Rock  
Your World & Some.  
Sexy Katerina - I am new to the Bay Area and ready to meet you.  
Please call me for happy time.  
Incall  
160 for 1 hour  
Call for appointment  
510-857-XXXX or 510-586-XXXX no private call.  
("Oakland/East Bay Escorts," Backpage 2012).*

If one ad advertises for two or more girls, this could indicate a shared management situation. This one is tougher for an algorithm to discern by itself because it could really be two girls who are friends working together or it could be two girls under one pimp; such a case would require the help of a human analyst to better differentiate, after it has been pinpointed by the automated detection algorithm.

#### *Shared Management Example Analysis*

One example of what could be a shared management situation took place on the Pittsburgh Backpage Escorts section. The group of posts was initially discovered from a manual review of the data. The similarity that all these posts share is that they advertise a girl and direct the buyer to call “Tony” or “Jen” (sometimes “Jenna”) to schedule an appointment so as to not interrupt the escort’s time. Using an automated search in Microsoft Excel, it was discovered that there were 278 such posts as of April 17<sup>th</sup> (with data collection beginning in October). All these posts came from the Pittsburgh Backpage Escorts section. In addition, in the whole set of posts, there were only *two* phone numbers that were repeatedly listed. Both of these phone numbers had a 412 (Pittsburgh) area code, suggesting that the operation was centered in or around Pittsburgh and had not moved. The posts also frequently reference “www.adorableplaymates.net”. The presence of an outside website suggests a business-framework of the organization.

Example posts from this situation are shown on the next page (with common phrases and phone numbers highlighted). Please note that these posts have been cleaned using our data preprocessing technique and therefore do not contain common filler words such as “the,” “a,” and “or.”

FULFILL_FANTASY_PLAYMATES_200HR_CASH_CREDIT_GORGEOUS_PROMPT_DISCREET_GIRLS_PICTURED_TIL_6AM_PICTURES_RECENT_REAL_OUTCALL_TEXT_PLEASE_412-306-XXXX_CHECK_LADIES_ADORABLEPLAYMATES_NET_TONY_JEN_TAKING_CALL_TIME_RUSHED_INTERUPPTED
HOTTIES_200HR_CASH_CREDIT_GIRLS_PICTURED_AVAILABLE_TIL_6AM_GORGEOUS_PROMPT_DISCREET_SERVICE_CHECK_LADIES_ADORABLEPLAYMATES_NET_OUTCALL_412-306-XXXX_TONY_JENNA_TAKING_CALL_TIME_TOGETHER_UNRUSHED_INTERUPPTED
BEAUTIFUL_GIRLS_OUTCALL_FAST_DISCREET_CASH_CREDIT_FEELING_FRISKY_WAITING_EXCITEMENT_EVENING_WARM_PICTURES_RECENT_REAL_BR_CALL_412-306-XXXX_TONY_JEN_TAKING_CALL_TIME_UNINTERUPPTED_OUTCALL_TEXT_FRIENDS_ADORABLEPLAYMATES_NET
HOTTIES_HIRE_CASH_CREDIT_GORGEOUS_PROMPT_DISCREET_TIME_LADIES_RUSHED_INTERUPPTED_GIRLS_PICTURED_REAL_RECENT_PICTURES_MULTI_GIRL_HOURS_CHECK_LADIES_ADORABLEPLAYMATES_NET_OUTCALL_TEXT_412-306-XXXX_CALL_TAKEN_TONY_JENNA_AFTERNOON_TIME_UNRUSHED_INTERUPPTED

Note: “outcall” is a situation in which the seller offers to come to the buyer’s house or location.

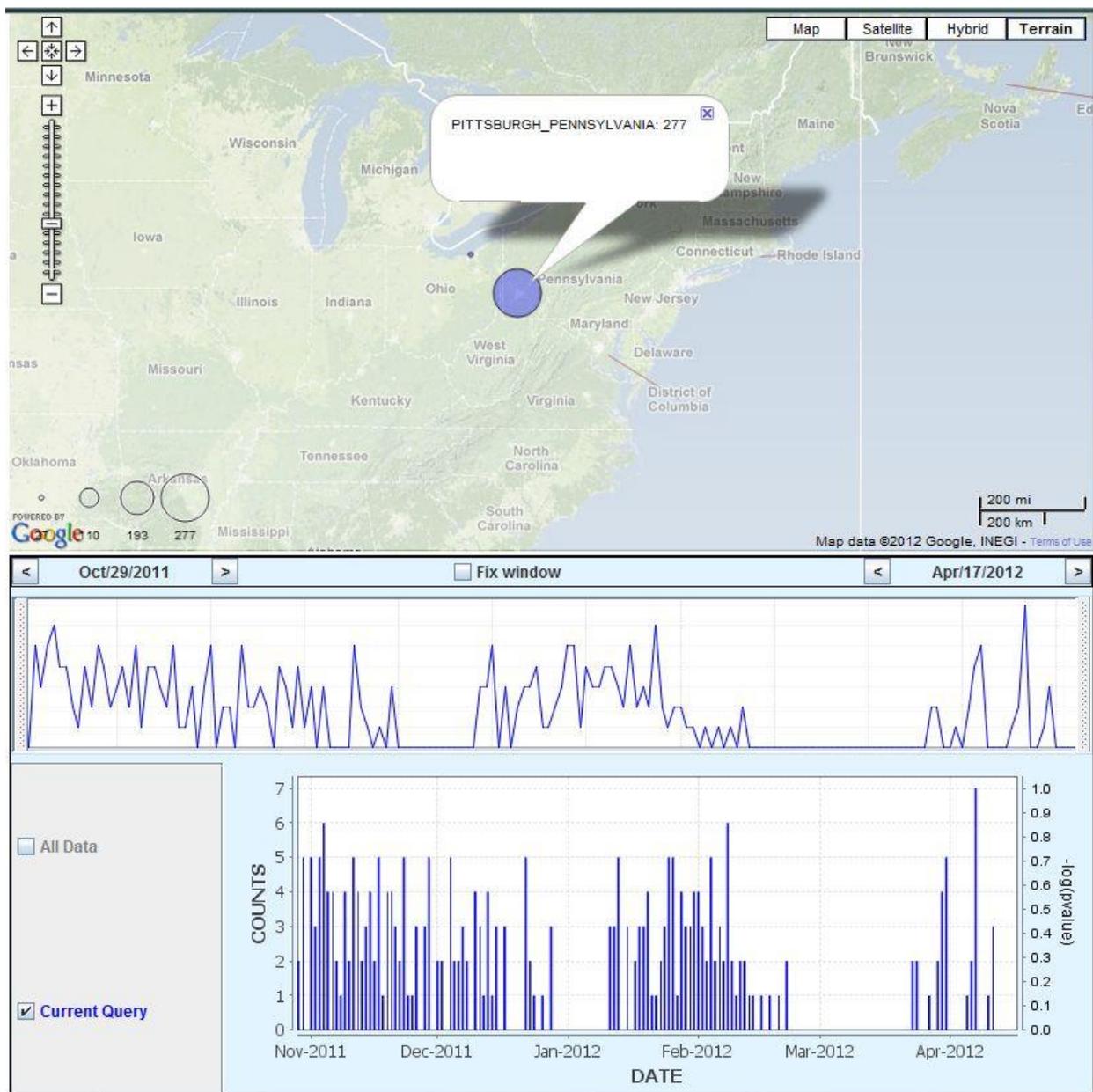
These posts exhibit many of the shared management triggers mentioned above. This example shows the importance of not only finding posts that are exactly the same, but finding posts that contain *similar language*. These posts would not have been identified as being written by the same author if the computer had only been directed to look for posts with exactly the same language. To summarize the evidence that points towards shared management in this situation:

1. Posts tend to only list one or two phone numbers in the whole group of posts.
  - a. The area code of the phone numbers corresponds to the area code of the city in which they were posted.
2. Posts each reference two people, one male and one female, who should be contacted to purchase time with the girl(s).
3. Posts offer multiple girls.
4. Posts frequently contain a reference to an outside website.

### Viewing the Data Spatially

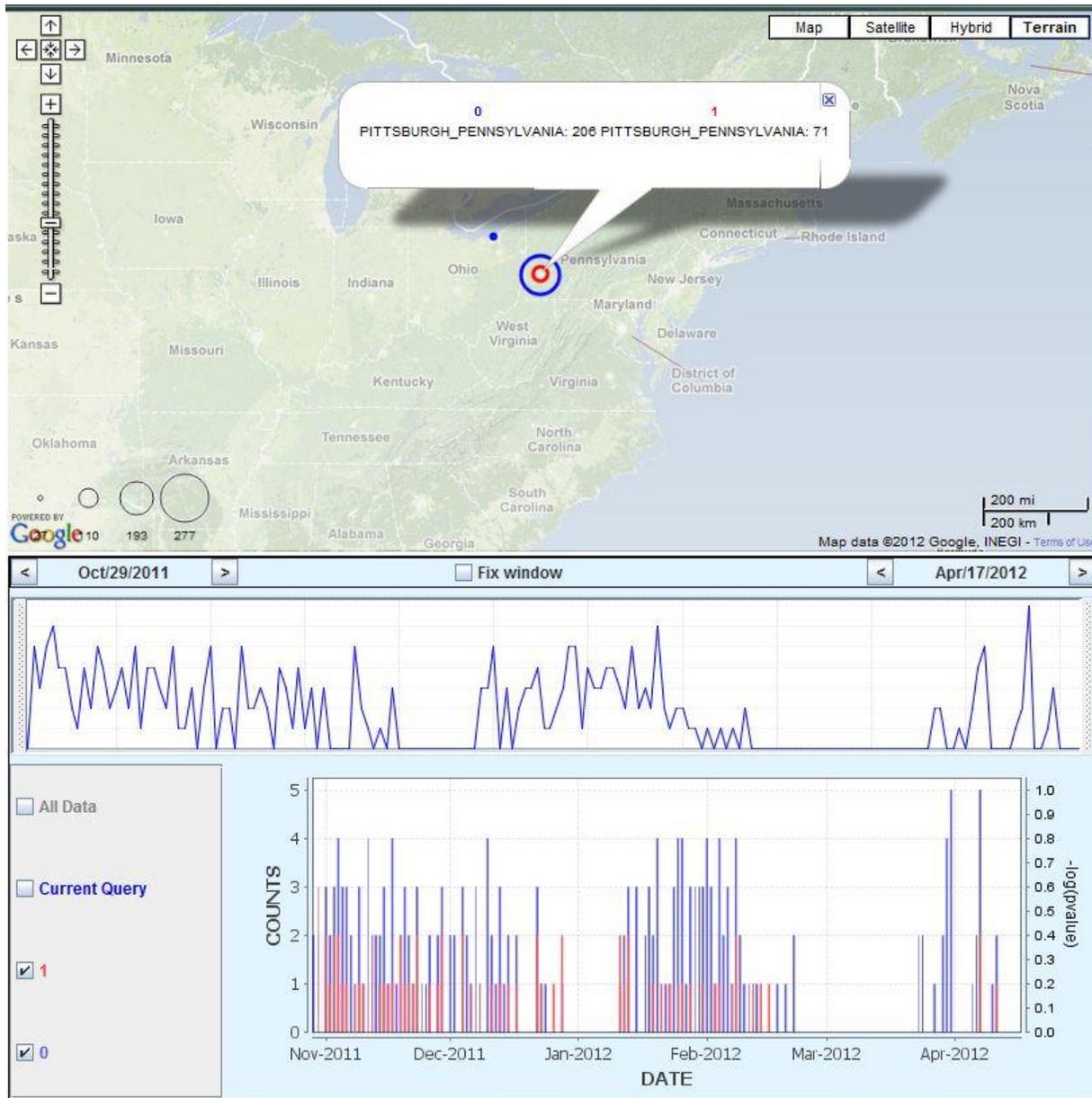
In order to view this data spatially in TCWI, I made a query on the data for the word “adorableplaymates.” That is, I made TCWI isolate only posts which contain that phrase. I did this because many of the posts with these 412 numbers also contain a reference to adorableplaymates.com. On the map below, one can see that all these isolated posts are posted in Pittsburgh, PA and that as of April 17<sup>th</sup>, there were 277 of these such posts in Pittsburgh, with one likely negligible outlier in Cleveland, OH. This means that there were a total of 278 posts

containing the phrase “adorableplaymates.” Below the map, the small graph shows post frequency over time. Note that the posts begin around the time that data collection began—that is, they were likely going on before this time. In addition, the posts continue sporadically up through the last period of data collection, with a small hiatus around March. Finally, the bar graph at the bottom shows the frequency of posts containing the word “adorableplaymates” over time with more specific labels on count and date.



The information in this graph is potentially valuable for a number of reasons. It allows the viewer to isolate posts on a large scale, view trends in post frequency, as well as view a spatial representation of the data. As will be shown in later examples, sorting and viewing the data this way may prove valuable for not only isolating and observing trends of posts, it may also be valuable in demonstrating movement of posts over time and space.

To see the distribution of posts containing one of the two phone numbers, the posts were split by dividing them into those posts which *did* contain Phone Number A (in red) and those which did *not* (in blue). The breakdown of this analysis can be seen on the following page. As one can see, 206 of the total 278 posts did *not* contain this phone number (0) and 71 did (1). A similar analysis can be done with the other phone number.



The information provided by this graph could be useful for law enforcement to identify the frequency that a group uses a phone number, or to identify whether a group uses two or more phone numbers simultaneously. For instance, if one phone number's frequency ended where another's began, that could indicate that the individual or group got rid of an old phone number and obtained a new one for security purposes.

Given this analysis, the Tony/Jen example could be a potential shared management situation. It is possible that the adjective “adorable” could suggest underage trafficking, but further investigation would be required in order to make further conclusions.

### *Identifying Geographic Movement*

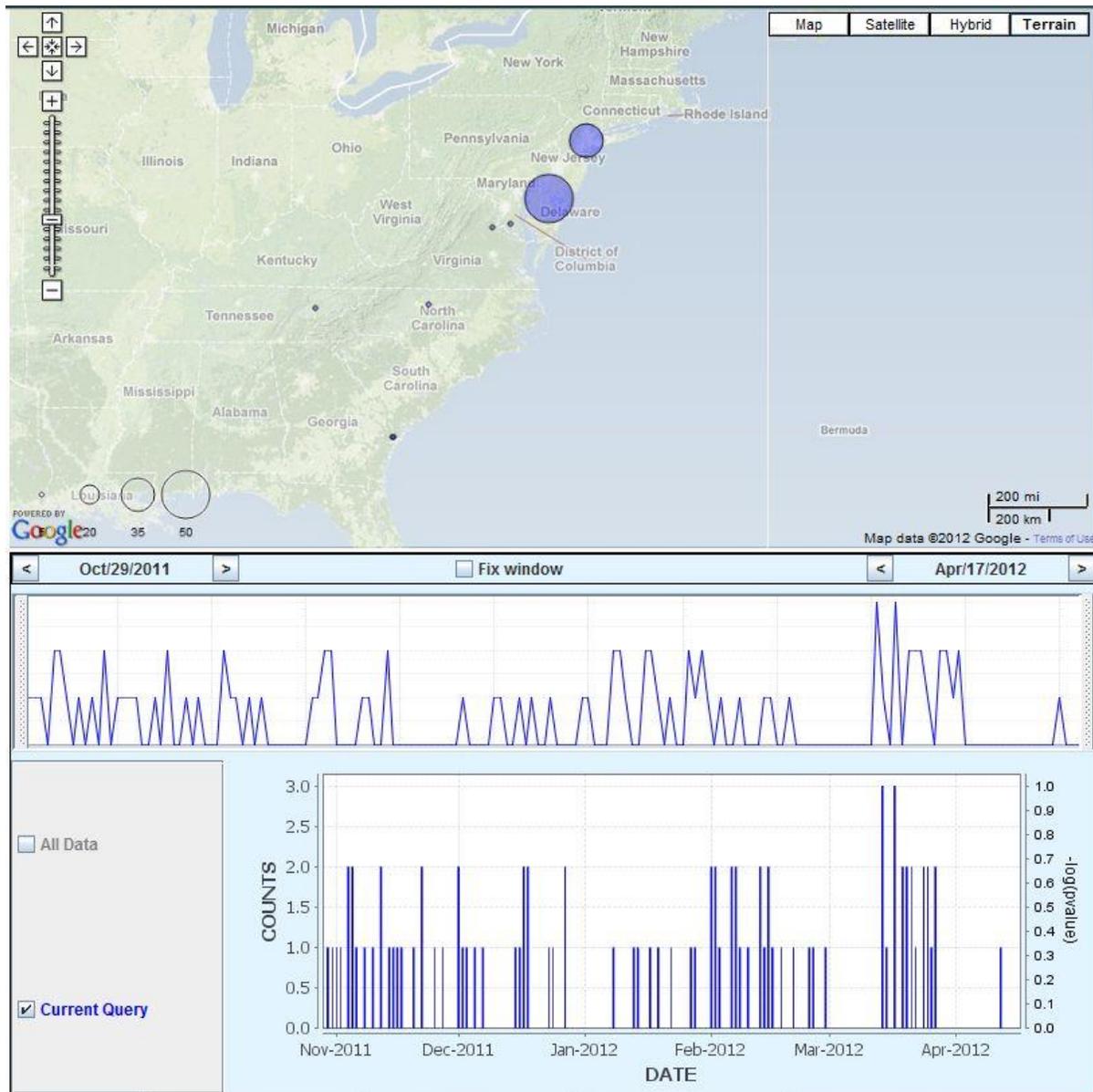
In addition to narrowing down posts that are more likely than others to be cases of underage prostitution or shared management, it is also crucial, if possible, to identify the movement of a post or author across cities and states in order to highlight a situation as a potential trafficking case. As mentioned earlier, this is further complicated by the fact that it is impossible to know whether or not an author is actually physically in or near the city in which he or she is posting. Even so, this analysis of the data will show that, in some cases, it is possible to track the movement of a post over time and geographic location. This can be done by tracking the movement of:

1. **The same exact ad**) across cities and states over time
2. **Similar ads** across cities and states over time
3. **The same phone number** across cities and states over time

### *Potential Movement Example Analysis*

One way to identify movement or spreading of posts is by isolating a group of posts, then separating them into groups by place and viewing this separation geographically. This was done for a group of posts containing a common phone number from the 502 area code. This number was chosen when, during a manual analysis of the data, it was noted that this number reappeared in a number of states. The query made on the data was to show all posts for which the value of containing that phone number is 1. That is, TCWI would display all posts that contain that phone number. The search yielded a total of 87 posts, spread across 7 states on the East Coast. The map below shows the distribution of the posts across the East Coast and the bottom

frequency graph indicates that the posts have been fairly consistent from the beginning of the data collection up until the data freeze on April 17<sup>th</sup>.



In order to view the movement of the posts over time, the subset of posts containing this phone number were split by location. That is, they were separated into categories based on their location, and then color-coded by location. Note that in the map and graphs above, 7 locations were identified, but in the following graphs, only the top 5 locations most frequently posted in were displayed (that is, the post in Tennessee and the post in North Carolina were removed) to

help graph visibility. The locations identified were Southern Maryland (red), Delaware (green), Fredericksburg, Virginia (orange), Central Jersey, New Jersey (light blue), and Savannah, Georgia (dark blue). In the graph below, you can see the distribution of the posts across cities over time. Most of the posts took place on the Delaware and Central Jersey Escorts pages.



To view the data more easily, the graph below shows the post frequency split by location and zoomed in. This view does not include the one post made in mid-April (which can be viewed in the graph above). Notable variations in location are identified with arrows. From left to right, they are: a post in Southern Maryland (indicated by “1”), a post in Fredericksburg, Virginia (2), two posts in Savannah, Georgia (3), and two more posts in Fredericksburg, Virginia (4). These diversions are notable because they are quite a distance from the two pages which were most frequently posted in, Delaware and Central Jersey.



The movement of these posts across the East Coast Backpage Escorts pages could be indicative of geographic movement of groups or individuals as well. This information could be useful for law enforcement to identify patterns of movement. It also has the potential for analysts to identify which cities are the most frequently traveled, and what common circuits tend to be. Law enforcement agents, social workers, and various researchers are aware of general trafficking patterns (Raymond and Hughes 2001). This research, combined with the previously mentioned methods of identifying potential cases of trafficking and underage prostitution, could be useful for adding weight to the study of trafficking patterns, in order to identify, follow, and even predict the movement of traffickers and their victims.

### **Future Research**

In future research, we hope to be able to not only display spikes in post frequency when they are anticipated by foreknowledge of big events such as the Super Bowl in the USC study. Our research also intends to use algorithms to *first* detect statistically-significant spikes in post frequency in real time—as compared to historical trends within the data or a sub-group within the data—and *then* look for explanations for these spikes. Research will be increasingly fruitful as it is not dependent on prior knowledge but instead can identify events that were, until now, unidentifiable and inexplicable, and then look for plausible explanations for them. In this way, we hope to discover new knowledge about the processes that control the dynamics of sex advertising and use it in the context of potential trafficking.

There is a number of directions future data analysis could take, but research to date on this subject suggests that three directions could be especially fruitful:

- 1) Photo analysis, first to enable matching of posts on the basis of similarity of the attached pictures, with a possible further extension towards identifying girls who are, for instance,

both advertised in escorts ads and identified as missing by the National Center for Missing and Exploited Children

- 2) Identifying posts that are probably posted by the same individual, whether or not the posts contain the exact same language. Relevant foundational research towards probabilistic matching of patterns is already underway at the Auton Lab in the Carnegie Mellon School of Computer Science and it could be leveraged in our future work.
- 3) Analyzing how the area codes of contact phone numbers listed in ads correspond to the area codes of the regions the ads are posted in. This could potentially lend insight into which regions of the country tend to have more frequent traffic. Regions with a high density (that is, most phone number area codes correspond with the region) could suggest areas of less frequent movement and regions with a lower density could suggest that those who tend to post in that region move more (because their area code corresponds less with the region)

Further research into these and similar questions are important because they have the potential to further an understanding of patterns of human trafficking via publicly available online classifieds websites.

## Appendix A

### Cases of Trafficking on Craigslist and Backpage

Note: the locations of trafficking (under “Where”) represent where news reports say the base of trafficking was located, but does not necessarily indicate the city or cities in which the girl(s) was trafficked.

#### Cases of Trafficking on Craigslist

When reported	Where	Situation	Where reported
May 3, 2010	Girl was trafficked across the East Coast	Three sisters (ages 11, 14, and 16) were sold for sex on Craigslist by a 32-year-old man	Rachel Lloyd, “An Open Letter to Jim Buckmaster,” Change.org.
Aug 10, 2010	Girl was trafficked in Los Angeles, Houston, Little Rock, Las Vegas	11 year old girl advertised via Craigslist by a 28-year-old man, under the name “Daddy Day Care”	Malika Saada Saar, “Craigslist’s shame: Child sex ads,” CNN.

#### Cases of Trafficking on Backpage

When reported	Where	Situation	Where reported
Aug 31, 2010	Girls trafficked in Nashville, TN	37-year old pimp trafficking two 17-year old girls (who were from Georgia and Virginia) for sex	Tim Wetzel, “Man, Two Teen Girls Arrested in Prostitution Sting,” News Channel 5, Nashville, TN.
Aug 19, 2010	Girl was trafficked in Denver, CO	Three men and two woman arrested for trafficking a 17-year old runaway from Colorado Springs	UPI, “6 charged with child prostitution,” United Press International, Inc. Felisa Cardona, “Six indicted on sex charged after teen’s kidnapping,” The Denver Post.
Nov 25, 2009	Girls were trafficked in Tea, SD	A man and woman may face life in prison for trafficking underage girls (ages and places of origin were undisclosed) under the name “South Dakota Escorts” on Backpage.com	Ben Dunsmoor, “Couple Faces Life in Prison for Sex Trafficking,” Kelo-TV, Sioux Falls, SD.

## Appendix B

### Email to Backpage.com

**Subject:** Research Question  
**From:** "Emily Claire Kennedy" <ekennedy@andrew.cmu.edu>  
**Date:** Wed, April 4, 2012 5:53 pm  
**To:** support@backpage.com  
**Priority:** High  
**Options:** [View Full Header](#) | [View Printable Version](#) | [Download this as a file](#) | [Add to Address Book](#) | [View Message Details](#)

Hello, I already submitted this email through the "Contact Us" page on the Backpage website, but I wanted to ensure that it got to the right channels.

To whom it may concern:

My name is Emily Kennedy and I am a senior studying at Carnegie Mellon University in Pittsburgh, Pennsylvania. I am doing research on how online media are used in the trafficking of human beings for sex. I am particularly interested in studying the posts in the "Escorts" section of Backpage to gain insight about the trafficking business in order to contribute to anti-trafficking research.

I know that preventing human trafficking—especially the trafficking of minors—is a goal shared by the founders of Backpage. My work is being done under the oversight of Dr. Artur Dubrawski of the CMU School of Computer Science and Dr. Jay Aronson of the CMU Department of History. Carnegie Mellon University is a non-profit accredited institution with a strong reputation in scientific research. The goal of my project is to contribute to the common good through a sociological study of this data.

I would like to obtain Backpage's permission to automatically gather the user-generated publicly available information in the "Escorts" section of the website via Backpage's RSS feed. This will not place a noticeable usage load on your site, and my data collection will not put any burden of time or effort on the Backpage personnel. Should I disseminate results of my research in the future, I will be happy to give Backpage due credit for letting me access and analyze the posts published by Backpage users.

Thank you for your consideration of this matter, and I look forward to hearing from you. I can be reached by email at [ekennedy@andrew.cmu.edu](mailto:ekennedy@andrew.cmu.edu).

Sincerely,  
Emily Kennedy

## Works Cited

- Aim Group. "Backpage replaces Craigslist as prostitution-ad leader." *Aim Group*. Advanced Interactive Media Group LLC, 19 Oct. 2010. Web. 24 Apr. 2012.
- . "Sites set combined record for online prostitution-ad revenue." *Aim Group*. Advanced Interactive Media Group LLC, 24 Feb. 2012. Web. 24 Apr. 2012.
- Backpage.com*. Village Voice Media. Web. 24 Apr. 2012.
- Cain Miller, Claire. "Craigslist Pulls 'Censored' Label from Sex Ads Area." *New York Times*, 9 Sept. 2010. Web. 9 May 2011.
- Chuang, Janie. "Beyond a Snapshot: Preventing Human Trafficking in the Global Economy." *Indiana Journal of Global Legal Studies* 13 (2006): 160. Web. 7 Apr. 2011.
- CNN. "CNN: Craigslist and the sex trade." Posted by "CNN" 3 Aug. 2010. YouTube video. 9 May 2011.
- Communications Decency Act of 1996, 47 U.S.C. § 230(c)(1), 1996.
- Craigslist.com*. Craigslist. Web. 24 Apr. 2012.
- Desplaces, Christian. "Police Run 'Prostitution' Sting; 19 Men Arrested, Charged in Fourth East Dallas Operation." *Dallas Morning News*, 21 Nov. 1992, 34A. Web. 24 Apr. 2012.
- Dickinson Goodman, Jessica and Matthew Holmes. "Can We Use RSS to Catch Rapists." Grace Hopper Celebration of Women in Computing, Portland, OR. Invited Conference Poster. Nov. 2011.
- Frolik, Cornelius. "Sex trade thrives by exploiting Internet: Tipp City case illustrates how Web is used to lure women into prostitution." *Dayton Daily News*, 27 Sept. 2011. Web. 24 Apr. 2012.
- Goodman, Mickey. "Sex Trafficking in the United States: Children Across America are Unseen Victims." *Huffington Post*, 23 Jan. 2011. Web. 9 May 2011.
- Hetter, Katia. "Fighting sex trafficking in hotels, one room at a time." *Cable News Network*. Turner Broadcasting System, Inc., 1 Mar. 2012. Web. 24 Apr. 2012.
- Interview with anonymous law enforcement agent. November 2011.
- Kloer, Amanda. "Los Angeles Pimp Arrested for Kidnapping, Trafficking Young Teens." *Change.org*. Change.org Inc., 4 Oct. 2010. Web. 9 May 2011.
- Kristof, Nicholas D. "How Pimps Use the Web to Sell Girls." *New York Times*, 25 Jan. 2012.

- Web. 24 Apr. 2012.
- . "Not Quite a Teen, Yet Sold for Sex." *New York Times*, 18 Apr. 2012. Web. 24 Apr. 2012.
- . "The Secret Owners Behind a Prostitution Website." *New York Times: Nicholas D. Kristof, On the Ground*, 31 Mar. 2012. Web. 24 Apr. 2012.
- . "Where Pimps Peddle Their Goods." *New York Times*, 17 Mar. 2012. Web. 24 Apr. 2011.
- Latonero, Mark, Genet Berhane, Ashley Hernandez, Tala Mohebi, and Lauren Movius. "Human Trafficking Online: The Role of Social Networking Sites and Online Classifieds." *Center on Communication Leadership & Policy: Research Series*. University of Southern California, September 2011, 26. Web. 24 Apr. 2012.
- Lloyd, Rachel. "An Open Letter to Jim Buckmaster." *Change.org*. Change.org Inc., 3 May 2010. Web. 24 Apr. 2012.
- Matyszczyk, Chris. "Craigslist removes 'censored' bar from site." *CNet News*. CBS Interactive, 8 Sept. 2010. Web. 24 Apr. 2012.
- Michels, Scott and Rich Esposito. "Attorney General Says Craigslist Agrees to End Erotic Ads." *ABC News*. ABC News Network, 13 May 2009. Web. 24 Apr. 2012.
- Polaris Project. "Domestic Sex Trafficking." *The Polaris Project*, n.d. Web. 9 May 2011.
- . "Myths and Misconceptions." *The Polaris Project*, n.d. Web. 24 Apr. 2012.
- Quantcast. "Backpage.com." *Quantcast*. Quantcast Corporation, n.d. Web. 24 Apr. 2012.
- Ray S., Michalska A., Sabhnani M., Dubrawski A., Baysek M., Chen L., and Ostlund J. "T-Cube Web Interface: A Tool for Immediate Visualization, Interactive Manipulation and Analysis of Large Sets of Multivariate Time Series." *American Medical Informatics Association (AMIA) Annual Symposium*, Washington, DC, 2008.
- Raymond, Janice G. and Donna M. Hughes. "Sex Trafficking of Women in the United States: International and Domestic Trends." *National Criminal Justice Reference Service*. U.S. Department of Justice, 17 Apr. 2001, 54. Web. 24 Apr. 2012.
- Sabhnani, Maheshkumar, Artur Dubrawski, and Andrew Moore. "T-Cube: Quick Response to Ad-Hoc Time Series Queries against Large Datasets." *Carnegie Mellon University Machine Learning Department*. The Auton Lab, n.d. Web. 24 Apr. 2012.
- Schapiro, Rich, Kerry Burke, and Tony Sclafani. "Hooker's New Hot Spots: Young Girls Lured by 'Oldest Profession.'" *Daily News (New York)*. 30 Oct. 2005, 14. Web. 24 Apr. 2012.

Stoeffel, Kat. "Village Voice Media Getting Down and Dirty with Escort Ads." *New York Observer*. 26 Apr. 2011. Web. 24 Apr. 2012.

Stone, Brad. "Sex Ads Seen Adding Revenue to Craigslist." *New York Times*, 25 Apr. 2010. Web. 9 May 2011.

Trafficking Victims Protection Act of 2000, 22 U.S.C. § 7102(8)(a), 2000.

Village Voice Media. "Statement to Nicholas Kristof RE Backpage January 24, 2012." *Village Voice Media*. Village Voice Media Holdings LLC, n.d. Web. 24 Apr. 2012.

Wright, David. "Craigslist: Site for Sex Slaves' Story Saves Girl's Life." *ABC News*. ABC News Network, 6 Sept. 2010. Web. 9 May 2011.