4-2019

# Analysis of 2018 Human Trafficking Location Data from skipthegames.com

Emily Peterson

# Senior Honors Project – Analysis of Selected 2018 Human Trafficking Location Data

## Analysis of 2018 Human Trafficking Location Data from skipthegames.com


**Honors student author:**

Emily Peterson

**Capstone Team Members:**

Nathan Relyea

David Vandewark

Samuel Symmes

Paul Hoang

**Industry Sponsor:**

Sergio Caltagirone, Global Emancipation Network

**Faculty Advisor:**

Dr. Andrew Nuxoll



**April 2019**

**Project Introduction**

Computer science students in the Shiley School of Engineering complete a capstone project during their senior year. The project discussed here is a result of a collaboration between the Global Emancipation Network (GEN), an industry nonprofit group, and the University of Portland. GEN works to combat human trafficking through analysis of data gathered from a variety of sources and this capstone project worked to look at a subset of data scraped from the escort website skipthegames.com to see if any important trends could be found which could later be applied to prevent human trafficking. The web postings on the website advertised potential trafficking victims for sex work, and these posts were analyzed to correlate users with shared identification information between posts. Itineraries of potential movement of trafficking victims between locations were created to highlight any major trafficking corridors which can show geographical areas for law enforcement focus and reveal patterns of trafficking victim travel that can later contribute to identifying and intervening in specific situations. This capstone project shows the potential of working with large datasets scraped from the web to gather information that can be used to combat human trafficking.

**Background**

The Global Emancipation Network has created a platform to store data related to human trafficking collected from a variety of sources and make this data available to law enforcement and other trusted organizations which work to combat human trafficking. Part of this data is metadata scraped from escort classification websites through which traffickers advertise to clients. This includes time of posting, location, phone numbers and emails. In this analysis, the metadata scraped from web postings is analyzed. The capstone project completed the goal of presenting new findings based on analysis of this web data.

**Project Overview**

This capstone project worked to identify interesting patterns found in location data scraped from skipthegames.com. To complete this data was first gathered, then posts were correlated based on common trafficking victim. The location movement patterns connected to each user were further analyzed by looking at statistically significant routes taken by all users, creating a heatmap of all locations of trafficking across all users, and looking for specific locations or routes that showed interesting conclusions about trafficking routes.

**Project Methodology**

Data was scraped from skipthegames.com, a known escort agency site, by the Global Emancipation Network collection algorithms. Skipthegames.com was selected for this analysis due to known good granularity for US cities. A year of data was determined a realistic goal for data processing and collection which would also provide an adequate sample size of data points. Towards this, posts created during 2018 which GEN had scraped data were analyzed. These posts were correlated based on the common characteristics of email and phone number. Posts containing no identification information (phone or email) were discarded. If multiple posts contained common email or phone numbers, these were combined to create a user 'itinerary', or a list of all the locations visited by a user and the corresponding timestamp.

Example itinerary for a user:

| **user X itinerary** | |
|---|---|
| **Location** | **Timestamp** |
| New York | 1/3/18 10:32 |
| Texas | 1/5/18 08:23 |

| | |
|---|---|
| Arizona | 1/5/18 07:22 |
| New York | 2/3/18 06:33 |
| California | 2/17/18 03:33 |
| New York | 2/23/18 08:33 |

This allowed 'routes' to be determined for each user. A route is a movement between two locations, location A and location B for a user where the timestamp for the post containing location A is earlier than the timestamp for the post containing location B and the user was not associated with any other location in a post with a timestamp in between A and B.

For example, the above 'user X itinerary' would yield the following routes:

Route 1: New York -> Texas

Route 2: Texas -> Arizona

Route 3: Arizona -> New York

Route 4: New York -> California

Route 5: California -> New York

These routes were calculated at the state level where state location was available for multiple correlated posts.

Significance for a given route was calculated based on total number of locations found in data points between all correlated users, and the expectation of random distribution of routes across these selected states. The probability of randomly getting a given route was calculated as follows:

$$probablity = \frac{occurences\ of\ route\ start\ location}{total\ of\ all\ locations} \times \frac{occurences\ of\ route\ end\ location}{total\ of\ all\ locations - occurences\ of\ start\ location}$$

A geometric distribution was then used to calculate significance level for each route. The analysis found several routes with a p value less than .01. Since the data set was large, numbers in the geometric distribution became too small to be accurately held in primitive types in C#. The team built a custom float object in scientific notation with a base and exponent. The geometric distribution function then added together the probability of getting every count lower than the actual count found in the data analysis and returned one minus this sum as the p value.

Additionally, further analysis was performed on the user itineraries. For each user itinerary, the following items were calculated:

- the count of repeat visits to a single location
- the number of round trips between two locations
- the percent of locations logged within 24 hours of a previous location
- An estimation of a likely “home base”

The percent of locations logged within 24 hours of another was calculated to examine in a simple way when posts with different locations might not correspond to movement if geographically distant locations were often logged within hours of each other.

A user’s potential “home base” is determined by evaluating the most common round-trip return location for a user which is the most common location that a user leaves and then returns to. Users that have three or more locations in their itinerary with over 80% of these locations logged within 24 hours of each other are not included in this data set as it seems unlikely that this quick location switching corresponds to actual movement. Home bases were compared across users to find common home bases between different users. For example if a user has trips logged (in order) from Location A -> B, B -> A, A -> C, C -> A, their itinerary will have noted two round trips starting from and returning to location A. Since

location A is the most common round trip location with a count of 2, location A is assigned as this user's potential home base.

Additionally, for each location logged, the percentage of users who have that location as their probable home base was calculated. Given a location *L*, the home base percentage is given by the following equation:

$$home\ base\ percentage = \frac{number\ of\ users\ indicating\ homebase\ in\ L}{total\ number\ of\ users\ who\ visit\ location\ L} \times 100$$

Most common repeat trip analysis is performed by determining which trip between cities or states was logged the most times for a user, and then comparing that most common trip to those of all other users. Again, users that have three or more locations in their itinerary with over 80% of these locations logged within 24 hours of each other are not included in this data set. Once a full list of the top most common repeat trips is formed, the percentage of those who make a given trip more often than any other trip compared to all of those who make that trip is calculated.

**Data limitations**

- Data is not guaranteed to be pulled or scraped evenly from all US cities.
- Posts ignored if location format was not formatted as "City, State, USA"
- Posts ignored if State name not written as full name or standard two letter representation (i.e. only 'Washington' or 'WA' accepted for Washington State)
- Area code 307 filtered out due to scraper error. Due to an exception in the scraper which caused certain image ids starting with 307 to be incorrectly classified as phone numbers. To correct this error, phone numbers starting with 307 were filtered out which also corresponds with a Wyoming area code.
- Common phone or email does not always correspond to movement for a victim or set of victims

- Date posted does not necessarily correspond with movement
- The 80% threshold for discarding location data where 80% of locations appear within 24 hours is only an estimation of a likely good threshold

Ultimately, we have no guarantee or verification that seeing multiple locations posted at different times corresponds to movement between those locations of a single person in the specified order. However, this analysis can show potential connections between different locations and related locations.

**Project Results**

A total of 1,179,871 posts were retrieved from skipthegames.com which resulted in 28,642 users. The distribution of total locations from posts which included a phone number and/or email is shown in a heat map in Figure 1 (below). These locations are latitude/longitude coordinates associated with a post.


*Figure 1: A heat map created using Bing Maps API showing locations from posts which included a phone number or email*

Out of these locations, 288 significant routes between states were found with all fifty states appearing. Of these routes, 110 (38%) were one-way routes. Note that for any of these one-way routes between a location A and a location B, the data could contain routes from B to A, but the route from B to A was not a significant route (had a p-value greater than .01).

These routes are represented in Figure 2 (below) where the route endpoints represent a state and are placed in the center of a state.


*Figure 2: A map created with Bing Maps API showing the significant 288 routes between states, each route has a random color*

Additionally, several interesting results from analyzing user itineraries were found:

Ogden, Utah

- 36.667% of the 60 users who post in Ogden, UT indicate that Ogden is their home-base
- Ogden is found twice in the top 5 most common city repeat trips by user, with Ogden to Salt Lake City being the most common trip for 12 users and Ogden to Provo being the most common trip for 11 users
- 66.667% of the 18 users who travel from Ogden to Salt Lake City make this trip more than any other trip

A specific address in Carlsbad, New Mexico (street address withheld from document for privacy)

- 176 users have posted using a specific address in Carlsbad, NM. 18 of these users indicate that this location is their home-base (10.227%)
- The specific address in Carlsbad, NM is found three times in the top 10 most common city repeat trips by user, with Odessa to the Carlsbad, NM address being the most common trip for 11 users, the Carlsbad, NM address to Lubbock, TX being the most common trip for 8 users, and the Carlsbad, NM address to Odessa being the most common trip for 7 users
- 28.947% of the 38 users who travel from Odessa to the Carlsbad, NM address make this trip more than any other trip

Other

- Only 0.513% of the 195 users who post in Hawaii indicate that Hawaii is their home-base
- Only 1.461% of the 2122 users who post in Florida indicate that Florida is their home-base
- Only 0.518% of the 193 users who post in Reno, NV indicate that Reno is their home-base
- 0.478% of the 209 users who post in Rochester, NY indicate that Rochester is their home-base
- Over 15% of the 368 users who post in Brooklyn or The Bronx indicate having home-base there.

**Project Tools**

Several tools were created during this project which were turned over to GEN at the completion for future use. This included the following:

- Data collection program:

  A program written in the coding language C# to download metadata from Splunk given valid credentials. This allows a GEN operative to plug in the correct username and password to a program and save specified data in the data format called JSON to a local computer or other host.

- Data reformatting scripts:

  Code written again in C# to take data which has either been directly downloaded from the Splunk web interface or pulled using the data collection program mentioned above and reformat this data into a format compatible with the tools which follow. By specifying different parameters in the code, data can be successfully put into the correct format.

- User correlation:

  After the data has been retrieved from Splunk and reformatted, this C# program will look through each post collected and group and posts with common phone or emails together. Common phone and/or email provide a good indicator of connection between two posts and the individuals associated with these posts. This user correlation program outputs a JSON formatted file which contains essentially a list of all users and the locations and timestamps associated with that user. It also contains the found phone numbers and emails for that user for potential further analysis. This user correlation is the necessary input for the following two tools on the list. An example user correlation object is shown below. In a user correlation output file, there will be many users.

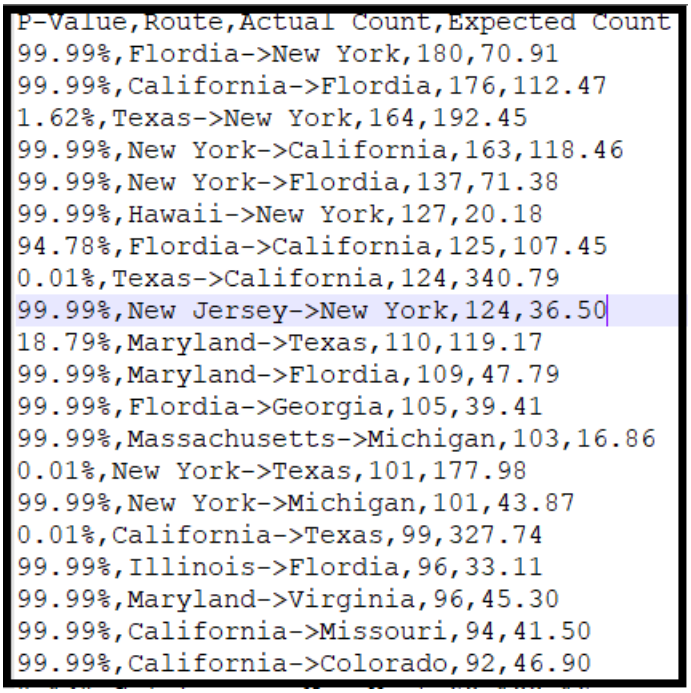
Example User Correlation Object:


```
{
  "user_id_11": {
    "collection_datetime": [
      "1547384040.000000",
      "1547325240.000000",
    ],
    "geo": [
      "Abilene, TX, USA",
      "Abilene, TX, USA",
    ],
    "geo_lat": [
      "32.4487364",
      "32.4487364",
    ],
    "geo_lon": [
      "-99.7331439",
      "-99.7331439",
    ],
    "phone": [
      "8888888888"
    ],
    "email": [
              "example@email.com",
              "anotherexample@email.com",
              "email3@email.com"
              ]
  }
},
```


Each user is given an identification number (11 in this example). The 'collection_datetime' parameter records the date of posting in a timestamp that corresponds to a year, month, day, second format. Three different types of location are recorded, in order a text based location (here two instances of 'Abilene, TX, USA'), the latitude and the longitude. The collection_datetime, geo, geo_lat and geo_lon parameters are all connected by index so that the first 'Abilene, TX, USA' corresponds with the first latitude and the first longitude as well as the first collection_datetime. The phone numbers and emails which were correlated with this user are also included.

- Statistical analysis of common routes between users:

  This C# script calculates the statistical significance of routes as discussed in the Project Methodology section. This script has modifiable parameters to specify the output including list of routes, count of these routes, significance and state versus city level routes. The output is a data file. A sample output of state level routes for a subset of data is shown below. This file can be used by further groups to analyze routes at state and city level. It can be modified or built upon depending on the research needs.



```
P-Value,Route,Actual Count,Expected Count
99.99%,Flordia->New York,180,70.91
99.99%,California->Flordia,176,112.47
1.62%,Texas->New York,164,192.45
99.99%,New York->California,163,118.46
99.99%,New York->Flordia,137,71.38
99.99%,Hawaii->New York,127,20.18
94.78%,Flordia->California,125,107.45
0.01%,Texas->California,124,340.79
99.99%,New Jersey->New York,124,36.50
18.79%,Maryland->Texas,110,119.17
99.99%,Maryland->Flordia,109,47.79
99.99%,Flordia->Georgia,105,39.41
99.99%,Massachusetts->Michigan,103,16.86
0.01%,New York->Texas,101,177.98
99.99%,New York->Michigan,101,43.87
0.01%,California->Texas,99,327.74
99.99%,Illinois->Flordia,96,33.11
99.99%,Maryland->Virginia,96,45.30
99.99%,California->Missouri,94,41.50
99.99%,California->Colorado,92,46.90
```

  In this sample, the significance of the route is listed first, the route is listed next in ‘First Location -> Second Location’ format, the actual observed count for the route is listed next and finally the expected count for that route if there was no significant to it (random) is listed.

- Further ‘Itinerary’ analysis of a given user:

  This C# script goes through a user correlation file and creates the ‘Itineraries’ and home base predictions discussed in the Project Methodology section. This output is also written to a file. A sample output is shown below. Again, by specifying different parameters, different output can be shown. The home base calculation takes place after the creation of the user itinerary object


```
{
  "id": "user_id_3938",
  "time": 1547279760,
  "location": "Berkeley",
  "visitcount": 3,
  "timeSincePrev": 277260,
  "repeatTripList": [
    "Vallejo",
    "Berkeley"
  ],
  "repeatTripCount": 2,
  "roundTripList": [
    "Berkeley",
    "Vallejo",
    "Berkeley"
  ]
},
```


  This sample user output includes the same style of identification number and timestamp for a user discussed in the user correlation section. This object shows information about ‘user_id_3938’ and ‘Berkley’ location showing a timestamp object timeSincePrev corresponding to how long since a post connected to this user had previously been published, a list of repeat trips and a list of route trips. A different portion of the Itinerary program would then be run to calculate home base and other statistics.

- Display of locations on map

  Using the Bing maps API, this html file takes an input list of locations and displays them on a map. Routes or data points can be displayed. This code created the maps shown in Figure 1 and Figure 2 and can be helpful to visualize data.

These created tools were the bulk of the work in this capstone project and are available to the Global Emancipation Network or future capstone teams that build off this work. This can allow for future researchers to continue this analysis with different data sources and modify and better the methodology described here.

**Conclusion**

Overall, the methodology and results described here provide optimistic insights for impacting human trafficking. Even with data limitations one web source was able to illuminate specific locations which could be investigated by law enforcement as well as provide a larger picture of where victims are being moved. While the specific location results of this analysis could be directly investigated, the biggest impact of this analysis is showing the potential of this type of work. Location data scraped from trafficking web posts has never before been analyzed in this way. This research shows that it is possible to correlate users across posts and observe significant movement. Continued work in this area could reveal pertinent patterns for law enforcement and agencies in the GEN network.

Future work in this area should include analysis on websites other than skipthegames.com and should look at areas outside the United States. Additionally, local analysis of a specific region using any available city data could show regionally interesting or significant routes that reveal more actionable information. It would also be interesting to examine how trafficking locations correlate with major interstates and highways.