GRIME: Graphing Related Illicit Massage Entities using Network Analysis

Sean Leader
Department of Statistics
California Polytechnic State University, San Luis Obispo
San Luis Obispo, USA
spleader@calpoly.edu

Bella White
Department of Computer Science and Software Engineering
California Polytechnic State University, San Luis Obispo
San Luis Obispo, USA
bwhite17calpoly.edu

Thea Yang
Department of Statistics
California Polytechnic State University, San Luis Obispo
San Luis Obispo, USA
tyang18@calpoly.edu

Amara Zabback
Department of Statistics
California Polytechnic State University, San Luis Obispo
San Luis Obispo, USA
azabback@calpoly.edu

***Abstract*—This paper discusses the network created in partnership with the Global Emancipation Network to identify potential criminal networks involved in Colorado's illicit massage industry. The network consists of publicly available data and can support in-depth investigations into suspected organized criminal activity involving individuals and businesses. In addition to this preliminary network, we developed a framework to reproduce and improve the network in other jurisdictions. This work serves not only as a useful tool for current investigations, but also as a proof of concept for future projects of similar nature in the fight against human trafficking.**

I. Introduction

Human trafficking is present and active in communities across the United States, with seemingly legitimate businesses serving as covers for forced sex work. One of the most common platforms to access forced sex work is through the services of illicit massage businesses. Illicit massage businesses refer to massage parlors that offer sex and prostitution services in addition to legitimate massages. According to the National Human Trafficking Hotline, massage businesses are the second most common form of sex trafficking [1], with estimates from 2018 suggesting that there are more than 9,000 illicit massage businesses active across the United States [2] generating $2.5 billion in profit per year.

Organized crime surrounding this industry often spans across more than one business, with criminals transferring victims between multiple locations to avoid suspicion. Therefore, it is important to disrupt the broader criminal networks rather than targeting individual sex workers or business owners as they can be victims too.

We conducted this research for the Global Emancipation Network (GEN) to help in the fight against human trafficking. GEN is a non-profit organization committed to disrupting and preventing human trafficking by working closely with researchers, policy makers, and law enforcement groups [3]. An important part of their operation is using modern technology to keep up with evolving trends in human trafficking and providing actionable data to lawmakers and law enforcement agencies.

Network analysis is a methodology with a broad range of applications, including analyzing social networks on social media platforms, optimizing company workflows, and discovering links between entities previously thought to be unrelated. Unlike other approaches, network analysis is focused on giving a holistic view of the interaction of entities on a broad scale.

The ultimate goal for this network is to uncover relationships and trends in the illicit massage business industry to support anti-human trafficking initiatives in Colorado and beyond. In this paper, we discuss our initial network product and results, examples of connections that can be revealed, and possible future work.

A. Ethical Considerations

We cannot discuss our methods and findings without first acknowledging that the network includes real people who may be victims of sex trafficking or another form of forced labor. Mishandling this data could result in harm to innocent people, which influenced our approach to the design process and analysis. Any relationships between individuals or entities to illicit massage businesses that exist in the network are not concrete evidence of involvement in illegal activity. For example, victims of human trafficking can appear on official business paperwork to obfuscate the involvement of the true owners. Additionally, a massage therapist could be a victim of sex trafficking or work at an illicit massage business with no knowledge of the illicit services offered. As a result, if the network is used or interpreted incorrectly, victims could be prosecuted in place of the actual perpetrators. Instead, the network should be used to uncover initial connections between entities and illicit massage businesses and should be further investigated by law enforcement agencies.

II. Project Purpose

This project's purpose is to create a well-documented framework for building a network of businesses and other entities in connection to illicit massage businesses. Given that human trafficking is a widespread issue not bound to any particular geographic region, it is paramount that this work can be applied beyond the scope of our particular network.

The three primary products we developed for GEN are as follows:

- An interactive network and initial exploratory results
- An adaptable network construction pipeline that may be modified for future uses
- Comprehensive documentation describing our process from start to finish

These deliverables will serve as both a benchmark for future networks and a guide to streamline construction of networks in different regions or for different purposes. The documentation for this project is particularly valuable as a majority of the work during this project was the network design phase. The documentation is intended to reduce time spent making basic design decisions in future efforts.

III. Data and Network Description

A. Network Scope

We targeted a geographic region with abundant and easily accessible data to reduce development complexity. Specifically, we used data from the state of Colorado due to its policies requiring massage therapist licensing as well as its easily accessible business data for the entire state. Further details on our project's scope considerations can be found in Appendix A.

B. Data Description

The network is entirely composed of data from publicly available sources. The majority of business data was sourced from the Colorado Secretary of State. Additionally, the network was supplemented by data from platforms such as Yelp and Rubmaps. Rubmaps functions as a review site where users seek out and share their experiences with illicit services at massage businesses.

For this project, any businesses listed on Rubmaps are labeled as ground-truth illicit massage businesses, and for the remaining extent of this paper referred to as an "IMB". A listing on Rubmaps does not confirm that illicit activities occur at such establishments, but we made this decision in order to establish a firm ground-truth within the network. This is not a perfect solution as legitimate massage parlors can also receive reviews on Rubmaps if customers are dissatisfied that they did not receive any illicit services. Filtering out these edge cases is discussed in Section V.

The network contains data of over two million records of all registered Colorado business entities from 1864 to present day. The database used to supply the network with this information is regularly updated and easily accessible, unlike the Rubmaps and Yelp data which requires complex web scrapes. As a result, business license data in the network is current as of May 2023, while data regarding Yelp and Rubmaps massage businesses is only current up to September 2021.

A comprehensive list of the data sources and the features included in each source can be found in Appendix B.

C. Feature Engineering

Some of our data sources included textual data that required additional extraction and analysis steps. We utilized natural language processing (NLP) techniques to extract names mentioned in the review datasets to increase the application of the data within the network. This extraction process resulted in a list of names which could represent either aliases or legal names of massage therapists associated with different massage establishments. For a detailed overview of the feature engineering methods we employed and their corresponding outcomes, please refer to Appendix D.

D. Making Connections

To create connections among entities, the network must be able to resolve format inconsistencies across multiple data sources. For example, because naming formats are variable across data sources, it is difficult to make direct entity name matches. Format variations include capitalization, punctuation, spelling, and different variables. To minimize the impact of these variations, we applied various manipulations to each of the initial data files to naming conventions conformed to a common format. This included capitalization, removal of punctuation, renaming of variables, and removing missing values. In addition to increasing connectivity of our network, these changes were also necessary to ensure that the data would be properly read into our network construction software, as it is sensitive to minor inconsistencies in formatting.

Unfortunately, data cleaning alone cannot remedy spelling inconsistencies. For instance, a business could be referred to as "Albert's Seesaws" on Yelp while being officially registered as "Albert Seesaw Inc." on the Colorado Information Marketplace. In cases where the business name is not identical across datasets, direct matches cannot be made with simple string comparisons, even if the business exists in both datasets. To reconcile such discrepancies, we used fuzzy matching. This approach, instead of seeking exact name matches, identifies names that are "suitably" similar. The specific degree of similarity had to be tuned to not be too restrictive while also not being too inclusive. Fuzzy matching allowed us to establish connections across various data sources, given that exact matches for most variables were exceedingly rare.

E. Network Description

The entities (nodes) of the network fall into several basic categories: businesses (such as IMBs, regular massage businesses, other businesses, and massage schools); business attributes (including phone number, address, and location); and people (massage therapists, reviewers, filing agents, and possible aliases).

Similarly, the relationships (edges) created between these nodes also fall into a few basic categories: location based (such

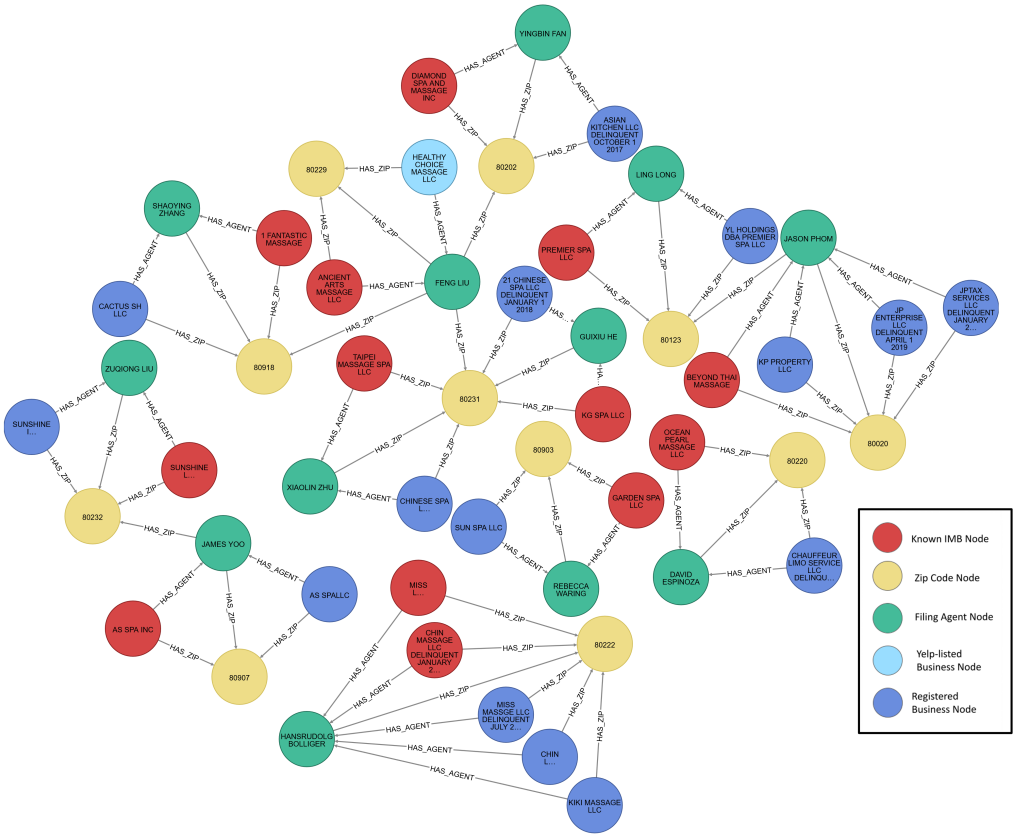

Fig. 1. Graphical result of businesses that share the same filing agent and zip code as an IMB. Cypher query can be found in Appendix E.

a business having a zip code or a business agent being located in a city); review-based (a reviewer reviewed a business or mentioned a massage therapist); person-based (businesses has an agent, therapist has an alias or works at a business); or miscellaneous (has a phone number).

The network that we produced encompasses a substantial scale, exceeding two million nodes and seven million relationships. With the inclusion of up-to-date data, we expect these numbers to grow steadily. The network is also highly adaptable, capable of being refactored to fit a diverse range of use cases. We designed it to be iteratively improved, where insights create novel inquiries that require further refactoring.

An exhaustive list and description of all nodes and relationships is included in the Appendix C.

IV. Results

A benefit of using a network to model entity relationships is the ability for interactive querying and filtering, offering profound insights into data patterns and relationships. Our network implementation was built using Neo4j, a graph database solution, and hosted on an Amazon EC2 instance. We used Cypher, Neo4j's native query language with syntax resembling other query languages like SQL, to construct the network, as well as read and analyze the final network.

Neo4j offers a range of flexible options to effectively work with the results obtained from a network query. These options include graph visualization tools that facilitate dynamic interaction with the graph as well as basic functionality to modify the network visualization for improved clarity and comprehension. Furthermore, Neo4j provides the capability to export query results in either a traditional table or JSON structure. Both export options allow for further post-processing of the results using other tools or languages. With many different options to work with the network and its results, users have the flexibility to analyze the query results and associated information in a manner that best suits their needs.

Even basic queries on the network allow us to perform insightful analyses that uncover complex relationships, examples of which are discussed in Section IV.

A. Exploratory Data Analysis

1) Investigation 1: This section discusses an investigation into businesses that have a common filing agent with an IMB. Although the network doesn't currently contain data linking businesses to their owners, it does contain information about the agent who filed the business license. However, an agent that filed for an IMB may not have direct involvement in any criminal activity associated with IMBs, and may potentially be a trafficking victim themself. However, using business agents as a linking attribute between businesses is valuable for the purposes of uncovering broader criminal networks.

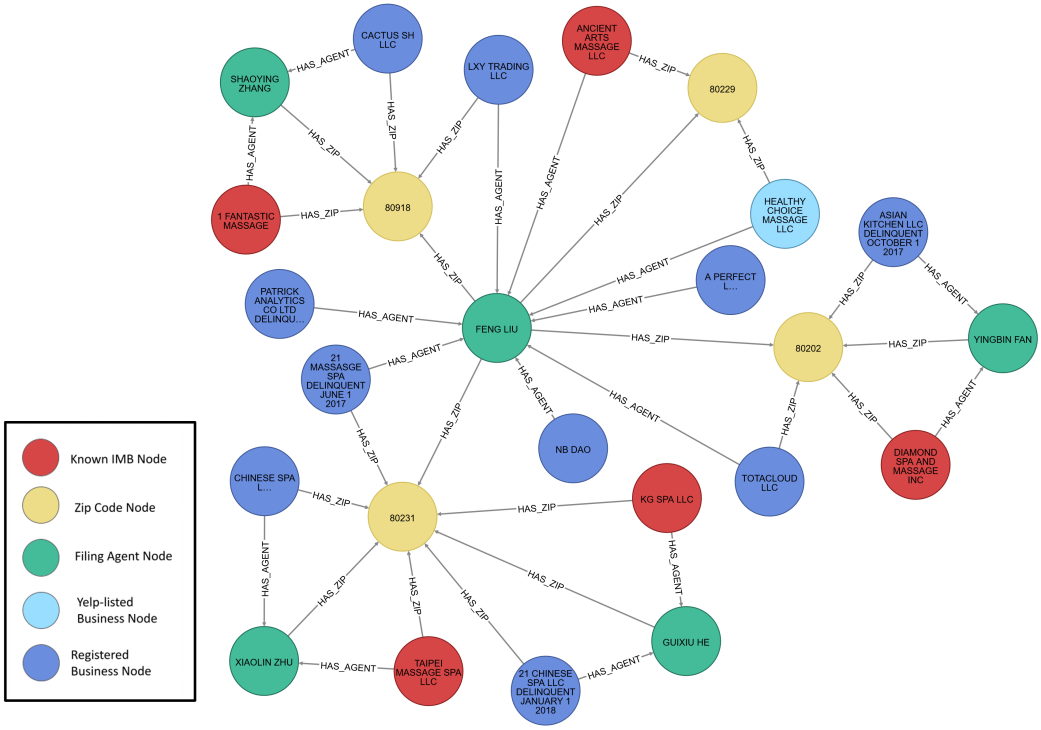

Fig. 2. Graphical result of businesses that share the same filing agent as an IMB, specifically focusing on the relationships between agent "Feng Liu" and other filed businesses. Cypher query can be found in Appendix E.

Fig. 1 presents the results of a network query that examines businesses sharing both the same zip code location and filing agent as an IMB. Visualizations such as Fig. 1 allow for a clear and intuitive overview of multiple relationships simultaneously. For instance, Fig. 1 reveals that the filing agent "Guixiu He" filed a license for both the IMB "KG Spa LLC " and another business registered as "21 Chinese Spa LLC". Because "21 Chinese Spa LLC" is not currently listed on Rubmaps, the network does not currently indicate that this business is an IMB. However, the indirect relationship between these two businesses may be reason for further investigation. The results of this query can be used as a starting point to investigate possible illicit massage businesses not listed on Rubmaps or confirmed to be illicit.

Visualizations like Fig. 1 demonstrate the relationships between entities. Neo4j allows you to aggregate and export data into a tabular file format to examine the properties of the nodes, such as business address or delinquency status. Table I is an example of such a tabular result, using a similar query as Fig. 1 to aggregate IMBs, businesses, and their corresponding addresses with shared filing agents. Table I is a subset of the complete tabular result, which can be found in Appendix E. This table reveals that some businesses filed by the same agent not only share the same zip code but also share identical street addresses with IMBs. For example, Table I shows that both "KG Spa LLC" and "21 Chinese Spa LLC" are registered at the address "1231 S PARKER RD," despite being registered as separate business entities. This example also demonstrates that conducting post-process analysis on specific node properties can aid in identifying potential suspicious connections to IMBs.

The network also provides the ability to leverage the results from one query to explore and further investigate specific relationships using another query. To demonstrate, Fig. 1 shows that filing agent "Feng Liu" has an edge connecting to zip code node "80231", which is a node that shares edges with other agents and businesses. However, this specific query output doesn't immediately provide any context for the relationship between "Feng Liu" and this particular zip code. Therefore, we ran a new query that builds upon the first, specifically focusing on the relationship path between "Feng Liu" and other businesses, without the specific restriction that businesses share connection with the same zip code.

Fig. 2 visualizes the results of this subsequent query. This result not only shows all the other businesses that "Feng Liu" has filed for, but also provides more context to "Feng Liu's" particular association with the zip code "80231". As seen in Fig. 2, "Feng Liu" registered a business under the name "21 Massage Spa LLC" but agent "Guixiu He" registered for a separate business under the name "21 Chinese Spa LLC" within the same zip code region. This result shows a similarity in both location and business name, but the separate registration of the business license by different filing agents. The patterns of nodes and edges as shown in this example can

TABLE I
SUBSET OF TABULAR-FORMAT RESULTS FOR BUSINESSES THAT SHARE THE SAME FILING AGENT AND ZIP CODE AS AN IMB. FULL TABLE CAN BE FOUND IN APPENDIX E.

agentName	IMBs	otherBiz	IMBAddress	bizAddress
DAVID ESPINOZA	[OCEAN PEARL MASSAGE LLC]	[CHAUFFEUR LIMO SERVICE LLC DELINQUENT JANUARY 1 2022]	[1452 POPLAR ST]	[1452 N POPLAR ST]
FENG LIU	[ANCIENT ARTS MASSAGE LLC]	[HEALTHY CHOICE MASSAGE LLC]	[7334 WASHINGTON ST]	[7334 WASHINGTON ST]
GUIXIU HE	[KG SPA LLC]	[21 CHINESE SPA LLC DELINQUENT JANUARY 1 2018]	[1201 S PARKER RD]	[1231 S PARKER RD]

indicate abnormal relationships between two businesses that may initially appear unrelated.

Again, we would like to emphasize that the results obtained from the network should not be the sole basis for accusing or prosecuting any entity or person in relation to illicit activities. Rather, this discussion focuses on the relationships between data points, which can be shared with relevant groups for more comprehensive investigations.

2) Investigation 2: This investigation explores the connections between IMBs, massage businesses listings on Yelp, and the associated Yelp users and reviews. The queries we wrote for this investigation are based on the hypothesis that individuals who patronize IMBs for sexual services may also frequent other massage establishments for similar purposes, even if these businesses are not listed on platforms like Rubmaps. Furthermore, the queries are rooted in the assumption that the majority of transactions at illicit massage businesses are between male clients and female massage therapists [4].

Fig. 3 is the result of a query that filters male Yelp users to those who gave a 4-star rating or higher to an IMB, and subsequently displays all the other massage businesses that the user has reviewed. While this query primarily focuses on Yelp reviews, the network identifies any business that was also reviewed on Rubmaps as an IMB. The cross-referencing and matching of business listings between both platforms is discussed in Section III.D. A list of all massage businesses in Fig. 3 that have not yet been identified as an IMB can be exported from Neo4j and used for further exploration using the tabular and JSON export options described in Section IV.A.1.

A key advantage of the network lies in its ability to be continuously updated to incorporate new knowledge and insights. We can enhance the network by focusing our attention to the aliases of the massage therapists employed at these businesses. In this particular investigation, we tagged all the other non-IMB massage businesses depicted in Fig. 3 as noteworthy because they have been reviewed by users who have previously favorably reviewed IMBs. As a result, we updated the network with a new edge that links the aliases associated with the noteworthy businesses to other massage establishments employing persons with the same name. These aliases were generated using the feature engineering techniques discussed in Section III.C.

Fig. 4 demonstrates the outcome of this update, including the newly established edge: "SHARES_NAME". This type of edge was not created at network instantiation, but rather was added in response to questions and interest that arose from analysis, demonstrating how the network can be refactored as needed. Fig. 4 is the result of a query that searches for aliases that share the same name and are associated with businesses located in the same area (zip code), along with the corresponding filing agent for those businesses.

Regarding the information displayed in Fig. 4, this query result reveals that "KG Spa LLC" and "Diamond Spa LLC" are not only located in the same general zip code area, but also have employees working there under the alias "Amy". It is possible that these two "Amy" nodes refer to completely unrelated individuals. However, because both businesses operate in the same area, this suggest the possibility that a single person is working for both establishments under the same alias. Ethically, we do not want to accuse either "Amy" or any mentioned businesses of engaging in human trafficking activities, nor do we seek to classify "Amy" as either a perpetrator or victim. The insights provided serve solely as an initial observation of abnormal or suspicious connections. It is essential to handle such information with care and involve appropriate enforcement agencies to proceed further.

Additionally, Fig. 4 includes an alias node labeled as "Yelp." This error originated from our feature extraction methods, discussed in Section III.C, inaccurately categorizing the word "Yelp" as a name. Due to false positives from feature extraction and other data cleaning procedures, some alias nodes are not actually aliases, but rather innocuous phrases. However, users interacting with the network can catch these inaccuracies when post-processing and analyzing query results. Reducing the amount of false positives in the name extraction process is an area that can be addressed in future work.

V. FUTURE WORK

The project of building a network to uncover connections between illicit massage businesses and related entities is far from complete. As a proof of concept, our work is the first step in what has the potential to be a fully comprehensive network supporting law enforcement agencies, human trafficking task forces, and research initiatives. The work can be expanded in any number of areas, such as data sources, feature engineering, or more insightful analytical techniques. Potential areas of future work are included in the following sections.

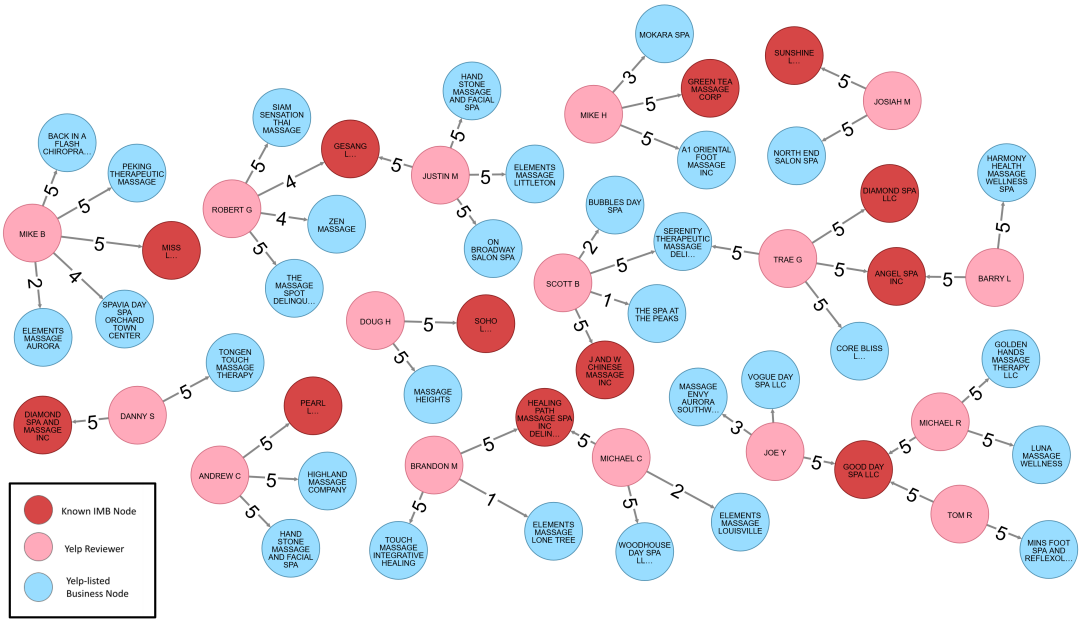

Fig. 3. Graphical result of massage businesses reviewed by Yelp users who have given an IMB a 4-star or higher review. Edges represent the numerical rating of the review. Cypher query can be found in the Appendix E.

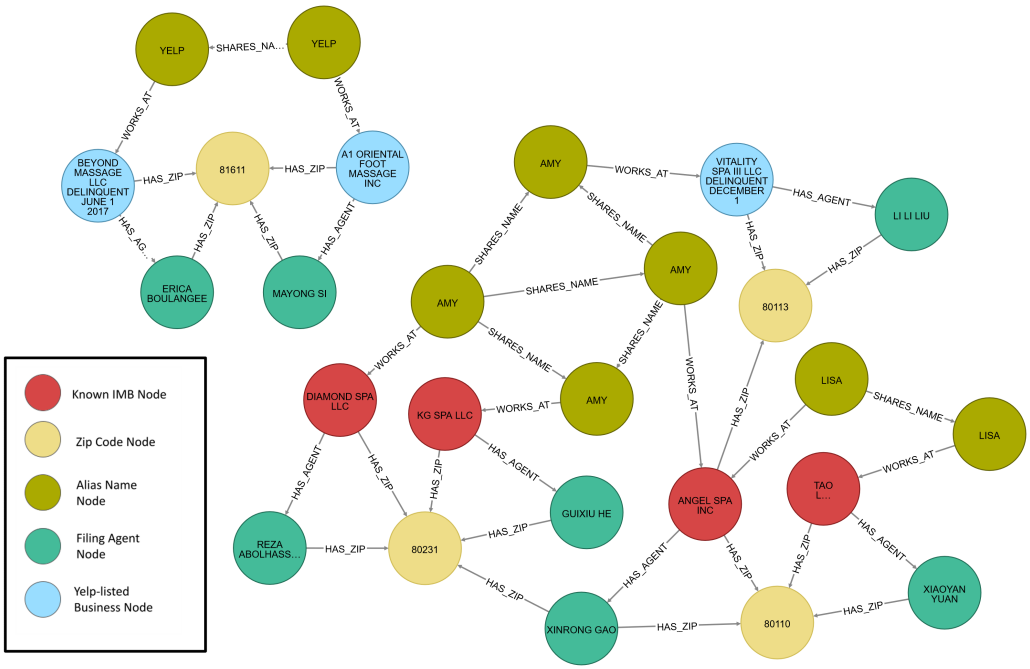

Fig. 4. Graphical result of aliases with the same name who work at businesses located in the same zip code area. Cypher query can be found in Appendix E.

A. Data Acquisition and Processing

As of this report, our network is not automatically updated when new data becomes available on the Colorado Information Marketplace. To address this, future work on this project should integrate data acquisition as part of the application pipeline so that the network is automatically updated. Furthermore, the current network implementation used easily-accessible and publicly available data for ease of network design and construction within the project timeframe. As a result, we did not use data sources that would require additional scraping or extraction steps. For example, the Colorado Information Marketplace also provides records of business name changes as well as documents pertaining to any disciplinary actions involving the business in .pdf format, which does not transcribe easily to a tabular format.

Additional avenues to explore can involve more extensive pre-processing techniques to extract further insights from existing data. The feature engineering conducted in our project represents only a fraction of the possibilities, particularly in regards to natural language processing (NLP) applied to Yelp and Rubmaps reviews. A particularly valuable application of NLP is sentiment analysis, which involves extracting the sentiment expressed in a review through relevant keywords and phrases. Implementing sentiment analysis can help filter out Rubmap reviews that reflect dissatisfaction with the absence of illicit services and make the ground-truth IMB labels in the network more precise.

B. Establishing Connectivity

Although our network has over seven million connections, not all possible connections between entities were successfully formed. One explanation for this stems from data cleaning, in which we deliberately removed observations with poor formatting or observations with insufficient information. Future work may seek to refine the data cleaning process to minimize the information lost while still maintaining consistent formatting. Other efforts to increase connectivity may introduce new data sources into the network, such as GIS technology which may help make more robust geographical connections between businesses, crime networks, or other locations of interest.

Connectivity of the network may also be improved in the refinement of fuzzy matching techniques used to join Yelp and Rubmaps business listings with their official Colorado business license registration. Improvements can be made by tuning how similar strings should be to be considered a match and using fuzzy matching to better inform the data cleaning process in order to prevent similar strings from being removed or altered. These improvements would optimize the ability of our pipeline to match records of the same business from different sources.

C. Integrating Graph Algorithms

An aspect of network analysis that we only briefly explored is leveraging the power of graph theory and algorithms. The work discussed in Section IV primarily relies on the exploratory abilities of the network, using query languages like Cypher. As a result, this exploratory analysis depends on the user to have some contextual knowledge of the current human trafficking landscape within the massage industry to accurately write and assess queries. The benefit of using graph algorithms is that one can programmatically evaluate the network's connectivity and communities because they are rooted in the mathematical properties of graphs, not underlying context.

There are many of these algorithmic tools in Neo4j's Graph Data Science library. In particular, we are interested in further explorations that use Louvain's Community Detection algorithm. This method evaluates how closely or densely related certain nodes are, and aggregates closely related nodes into a community [5]. This method can be beneficial for our network analysis because it can programmatically reveal the communities of businesses and other entities that may be related to IMBs, as well as justify its findings with the appropriate metrics. Applying this and other graph algorithms to the network would automatically produce more empirical results that can be passed along to interested investigative groups.

VI. Conclusion

We created a network to identify organized illicit massage business enterprises in Colorado, and included comprehensive documentation that allows this network to be reproduced and improved upon by the Global Emancipation Network in their continued efforts to combat human trafficking. The network serves as a framework for future networks in other geographic regions with many avenues and possibilities to enhance and expand upon the network. The presence of real victims in our data challenged us to consider the ethics of every decision we made throughout the design process to ensure we did not cause any additional harm. This network is meant to help victims of sex trafficking, not target them or accuse them of criminal wrongdoing. Thus, the end goal of the network is to provide law enforcement agencies with actionable information that will allow them to investigate potential criminal organizations, not individuals, involved in sex trafficking. We hope that using networks to identify the presence of organized crime will make a lasting and meaningful contribution to the fight against human trafficking.

VII. Acknowledgements

We would like to thank Sherrie Caltagirone, Margaret Tobey, and Joe Percivall from the Global Emancipation Network for their guidance and support through this project. We would also like to thank our professors and advisors, Drs. Alex Dehktyar and Hunter Glanz, for their education and mentorship. Many thanks to everyone who supported us as we created a network that we take pride in.

References

[1] "National Statistics." National Human Trafficking Hotline, 15 Nov. 2022, https://humantraffickinghotline.org/en/statistics

[2] "Hidden in Plain Sight - Polaris Project." *Polaris Project*, Apr. 2018, https://polarisproject.org/wp-content/uploads/2018/04/How-Corporate-Secrecy-Facilitates-Human-Trafficking-in-Illicit-Massage-Parlors.pdf.

[3] "The Global Emancipation Network Approach, Mission, and Offerings." Global Emancipation Network, 1 Sept. 2020, https://www.globalemancipation.ngo/global-emancipation-network-mission-offerings/.
[4] Brannon, Robert. "Prostitution: Key Facts and Analysis, in Brief - National Organization for Men Against Sexism." National Organization for Men Against Sexism, 22 Feb. 2022, nomas.org/prostitution-key-facts-and-analysis-in-brief.
[5] Gupta, Mehul. "Community Detection in a Graph Using Louvain Algorithm With Example." Medium, 13 June 2023, medium.com/data-science-in-your-pocket/community-detection-in-a-graph-using-louvain-algorithm-with-example-7a77e5e4b079.
[6] "Human Trafficking Statistics by State 2023," World Population Review, https://worldpopulationreview.com/state-rankings/human-trafficking-statistics-by-state.
[7] "Colorado Human Trafficking Council – Laws and Legislation." Colorado Human Trafficking Council – Laws & Legislation, sites.google.com/state.co.us/human-trafficking-council/human-trafficking-resources/laws-legislation.
[8] "EntityRecognizer · spaCy API Documentation." EntityRecognizer, spacy.io/api/entityrecognizer.
[9] "Tagger · spaCy API Documentation." Tagger, spacy.io/api/tagger.
[10] "The Stanford Natural Language Processing Group." The Stanford Natural Language Processing Group, nlp.stanford.edu/software/CRF-NER.shtml.

APPENDIX

A. Data Scope Decisions

When determining the specific geographic scope for this project, we took into account three key factors: demand, data availability, and data completeness. Based on data already collected by GEN at the onset of this work, we ultimately narrowed the scope to two options: California and Colorado.

California emerged as an appealing choice for our project due to its high incidence of human and sex trafficking, surpassing that of any other state [1], [6]. However, we encountered a challenging issue stemming from the state's legal framework, which lacks consistent restrictions for massage businesses. Instead, regulatory oversight is delegated to county and city jurisdictions. This would have required data acquisition and exploration for each specific geographic region—an arduous and time-consuming task. Furthermore, California does not mandate licensing for massage therapists, making it difficult to track and identify practitioners. Considering these factors, we concluded that focusing on the state of California would be impractical given our limited time and resources.

Conversely, Colorado stood out as a more viable option. The state has implemented centralized and comprehensive regulations governing massage therapy and human trafficking [7]. We were able to access highly complete datasets encompassing business licensing, massage therapist licensing, and information on delinquency statuses. Moreover, Colorado benefits from established task forces that could readily leverage the insights provided by our network analysis tool. Hence, we selected the state of Colorado as our primary subject for our network analysis and subsequent examination.

B. Data Sources

Data File	Source
Business Entities	Colorado Information Marketplace: https://data.colorado.gov/Business/Business-Entities-in-Colorado/4ykn-tg5h , downloaded April 2023
Massage Schools	Sourced from GEN's internal data library
Massage Therapists Licenses	Colorado Department of Regulatory Agencies: https://apps.colorado.gov/DORA/licensing/Lookup/GenerateRoster.asp ,downloaded April 2023
Rubmaps Reviews	Sourced from GEN's internal data library, scraped September 2021
Rubmaps Businesses	Sourced from GEN's internal data library, scraped September 2021
Yelp Reviews	Sourced from GEN's internal data library, scraped September 2021
Yelp Businesses	Sourced from GEN's internal data library, scraped September 2021

C. Network Schema

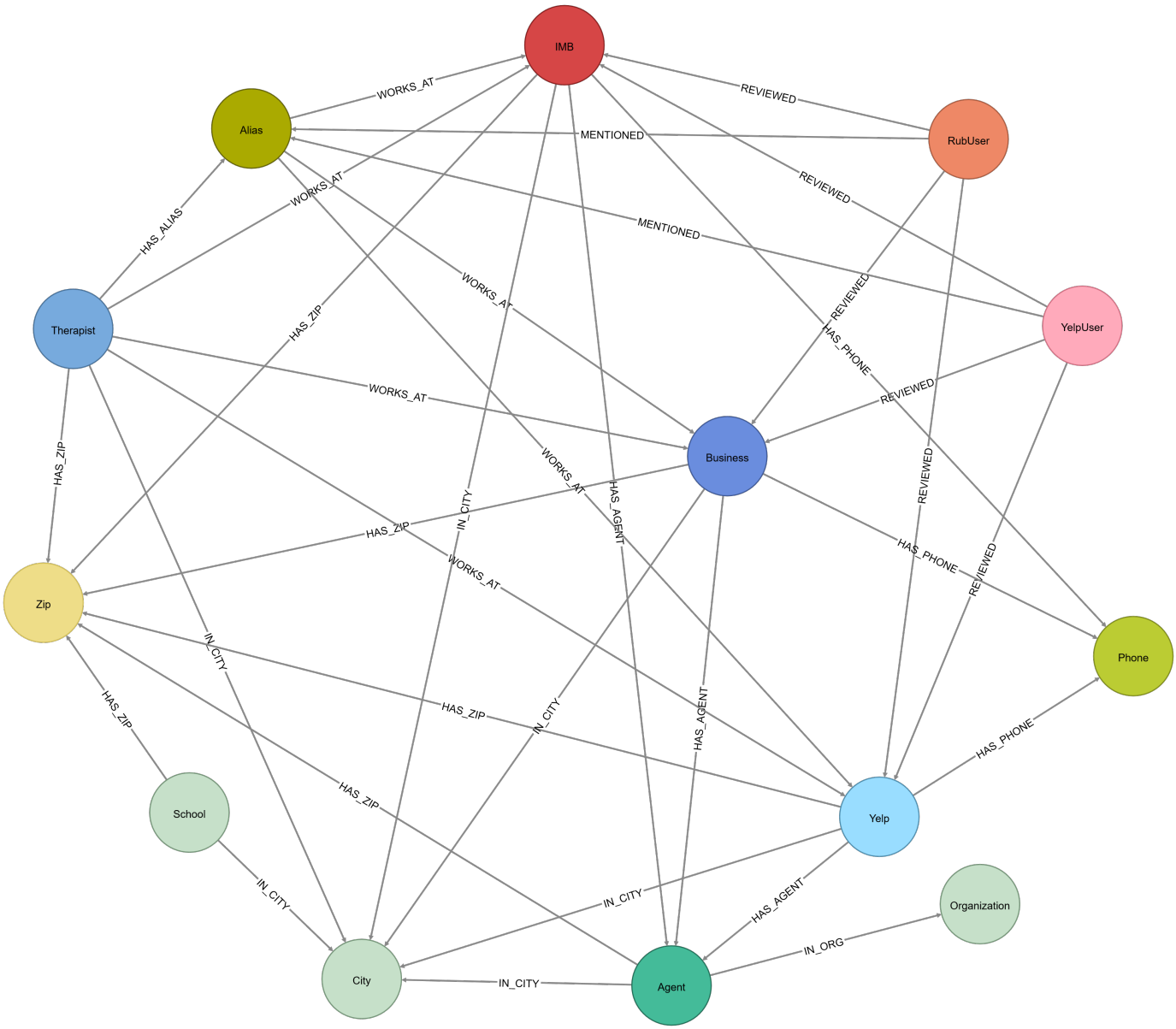

Fig. 5. Graphical representation of all possible nodes and edges included in the network.

Entity Type	Included in Graph
Nodes	: Massage Business : Illicit Massage Business : Business (other) : Massage School : Phone Number : Zip Code : City : Business Filing Agent : Yelp Reviewer : RubMaps Reviewer : Massage Therapist : Alias : Agent Organization
Relationships	**Location-based:** • HAS_ZIP • IN_CITY **Review-based:** • MENTIONED • REVIEWED **Person-based:** • HAS_AGENT • HAS_ALIAS • WORKS_AT • IN_ORG **Other:** • HAS_PHONE

Fig. 6. All possible nodes, corresponding colors, and possible edges/relationships. Which edges correspond to which nodes are shown in Figure 1.

D. Feature Engineering

Name Extraction

Name Extraction, also called Named Entity Recognition, is a common Natural Language Processing (NLP) tool that has a variety of applications. While each methodology offered differs in implementation details, the general process of NER and extraction is as follows:

1) Parts Of Speech tagging (POS tagging) to label each word (a.k.a. token, term) in a sentence with its part of speech, such as verb, adjective, noun, proper noun, etc.
2) Analysis of nouns and proper nouns in the context of the sentence structure for the purpose of labeling the type of entity. For example, a noun or proper noun could be a country, business, person, organization, etc.
3) Return of the specified entity. In this case, a person.

Many different open-source python packages exist to perform this type of recognition and extraction. A thorough investigation was performed to determine which package would work best for our specific needs, focusing the large corpus of data to be analyzed and efficiency when analyzing and extracting. Two packages revealed themselves as front-runners for our use: SpaCy [8], [10] and Stanford NER [10].

For our application of these methods in our corpus, we found SpaCy to be far more computationally efficient than Stanford NER, with limited discrepancies in accuracy. Given that SpaCy only missed a relatively small amount of the names identified by the authors of this paper and Stanford NER, and erred more on the side of false positives, SpaCy was chosen as the optimal method for this project.

Fuzzy Matching

Fuzzy matching, also known as approximate string searching and fuzzy string searching, refers to an NLP technique that matches words, phrases, or sentences that are not exactly the same, but are ”close enough,” defined by an edit-distance metric. The edit-distance metric used in our methodology was the Levenshtein Distance, which provides a number for how far the two strings are from being identical. This technique is especially applicable in this project, as many of our resources have data that refer to the same entity (person or business) but have slight variations in how they are recorded. For example, a business may be labeled as ”Business A” on Yelp or Rubmaps, but ”Business A LLC” in their licensing documentation. By fuzzy matching on both name of business and address, we were able to link several businesses to their Yelp reviews, which would have otherwise gone unmatched, meaning we have more accurate and thorough connections than we would have had, had we not applied these techniques.

Further refinement can be performed to ensure that the maximum amount of data is linked, such as tuning of the minimum Levenshtein distance for a match, better pre-processing of the strings, and a general improvement to the pipeline that processes the data to allow for smooth addition of future feature engineering.

E. Cypher Queries for Investigations and Additional Results

Cypher Query for Fig. 1:

```
match p=(i:IMB)-[:HAS_AGENT]-(a:Agent)-[:HAS_AGENT]-(b:Business)
match p2 = (i)-[:HAS_ZIP]-(z:Zip)-[:HAS_ZIP]-(b)
return p, p2
```

Cypher Query for Fig. 2:

```
match p=(i:IMB)-[:HAS_AGENT]-(a:Agent)-[:HAS_AGENT]-(b:Business)
match p2 = (i)-[:HAS_ZIP]-(z:Zip)-[:HAS_ZIP]-(b)
match p3 = (:Agent {iden:'FENG LIU'})-[:HAS_AGENT]-(:Business)
return p, p2, p3
```

Cypher Query for Table I:

```
match (i:IMB)-[:HAS_AGENT]-(a:Agent)-[:HAS_AGENT]-(b:Business)
match (i)-[:HAS_ZIP]-(z:Zip)-[:HAS_ZIP]-(b)
return a.iden as agentName, collect(i.name) as IMBs, collect(b.name) as otherBiz,
collect(i.address) as IMBAddress, collect(b.address) as bizAddress
```

Full tabular result for Table I:

agentName	IMBs	otherBiz	IMBAddress	bizAddress
DAVID ESPINOZA	[OCEAN PEARL MASSAGE LLC]	[CHAUFFEUR LIMO SERVICE LLC DELINQUENT JANUARY 1 2022]	[1452 POPLAR ST]	[1452 N POPLAR ST]
FENG LIU	[ANCIENT ARTS MASSAGE LLC]	[HEALTHY CHOICE MASSAGE LLC]	[7334 WASHINGTON ST]	[7334 WASHINGTON ST]
GUIXIU HE	[KG SPA LLC]	[21 CHINESE SPA LLC DELINQUENT JANUARY 1 2018]	[1201 S PARKER RD]	[1231 S PARKER RD]
HANSRUDOLG BOLLIGER	[CHIN MASSAGE LLC DELINQUENT JANUARY 1 2015,MISS MASSAGE LLC]	[KIKI MASSAGE LLC,CHIN MASSAGE LLC,MISS MASSAGE LLC,MISS MASSGE LLC DELINQUENT JULY 1 2018,CHIN MASSAGE LLC DELINQUENT JANUARY 1 2015]	[2640 S COLORADO BLVD,2242 S ALBION ST]	[2248 S ALBION ST,2640 S COLORADO BLVD,2242 S ALBION ST]
JAMES YOO	[AS SPA INC]	[AS SPALLC]	[828 E FILLMORE ST STE B]	[828 E FILLMORE ST]
JASON PHOM	[BEYOND THAI MASSAGE]	[JP ENTERPRISE LLC DELINQUENT APRIL 1 2019,KP PROPERTY LLC,JPTAX SERVICES LLC DELINQUENT JANUARY 1 2022]	[88 LAMAR ST]	[6143 W 113TH AVE,1704 ASPEN STREET,1704 ASPEN ST]
LING LONG	[PREMIER SPA LLC]	[YL HOLDINGS DBA PREMIER SPA LLC]	[8966 W BOWLES AVE]	[8966 W BOWLES AVE]
REBECCA WARING	[GARDEN SPA LLC]	[SUN SPA LLC]	[1107 S NEVADA AVE STE 113]	[1107 S NEVADA AVE SUITE 113]
SHAOYING ZHANG	[1 FANTASTIC MASSAGE]	[CACTUS SH LLC]	[6529 N ACADEMY BLVD]	[6529 N ACADEMY BLVD]
XIAOLIN ZHU	[TAIPEI MASSAGE SPA LLC]	[CHINESE SPA MASSAGE LLC]	[10011 E HAMPDEN AVE]	[8801 E HAMPDEN AVE]
YINGBIN FAN	[DIAMOND SPA AND MASSAGE INC]	[ASIAN KITCHEN LLC DELINQUENT OCTOBER 1 2017]	[1520 STOUT ST STE 3A]	[1520 STOUT STREET 3A]
ZUQIONG LIU	[SUNSHINE MASSAGE LLC]	[SUNSHINE MASSAGE INC]	[1864 S WADSWORTH BLVD 9]	[1864 S WARDSWORTH BLVD STE 9]

Cypher Query for Fig. 3:

```
match p= (y2:Yelp)-[:REVIEWED]-(yu:YelpUser)-[r:REVIEWED]-(:IMB)
where yu.gender = 'MALE' and toInteger(r.rating) >= 4
return p
```

Cypher Queries for Fig. 4:

```
match (y2:Yelp)-[:REVIEWED]-(yu:YelpUser)-[r:REVIEWED]-(:IMB)
where yu.gender = 'MALE' and toInteger(r.rating) >= 4
with y2
match p=(a:Alias)-[:WORKS_AT]-(y2)
match p2=(a2:Alias)-[:WORKS_AT]-(y:Yelp)
where a.name = a2.name and y2.name <> y.name
merge (a)-[:SHARES_NAME]-(a2)

match p2=(b2:Business)-[:WORKS_AT]-(a)-[:SHARES_NAME]-(a2)-[:WORKS_AT]-(b:Business)
with b, b2,p2
match p=(b)-[:HAS_ZIP]-(:Zip)-[:HAS_ZIP]-(b2)
return p, p2
```