Graph Matching on the Patterns of Life

Heavy Use of Facial Recognition Technology on China's Citizens, Government, and Industry

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Your Major

By

Bryan Kim

Fall 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Sean Ferguson, Department of Engineering and Society

Rosanne Vrugtman PhD, Computer Science

Daniel Graham, Computer Science

Introduction

Big data has evolved tremendously over the past decade and is still growing at a fast pace. The more intricate and complex the technology that comes up in the field becomes, however, more specific and even personal the data required is needed. This can be seen with the technical project from my internship last summer where the goal was to create a prototype pipeline for a client that would match differing phone numbers of the same user together. At the expense of using the call and messaging log data from such phone numbers extracted and put in a graph network, the created pipeline will then take the graphs and perform a matching analysis to indicate if any two numbers are close enough to be the same person. The purpose for this pipeline is to assist the Law Enforcement client in cases where suspects could switch phone numbers, saving time in investigating the identity of such phone numbers.. The STS topic, being loosely related to the technical piece, covers the impact of the heavy use of facial recognition technology and its use of personal user data on the ecosystem between citizens, government, and industry in China. The technical topic shows how detailed personal data is vital for various technologies to function correctly and reminds that the more data results in more analysis to extract. On the other hand, the STS topic on China gives a perspective of the over-use of such personal data and technologies and the effects it has on culture while informing others what to expect.

Technical Topic

Abstract

Though criminals constantly change their phone numbers, making it difficult for law enforcement to trace them, they rarely change their contacts. The process of re-identification can assist law enforcement efforts to trace the identities of unknown individuals by using graph networks from the interactions between the old and new phone numbers. The prototype pipeline I designed during my internship can extract call and message records from recorded logs and convert them to a graph network, creating a way to identify similar networks. The pipeline was tested using data found from Kaggle as well as some sample data that represented the client data. The outcome from the tests shows a promising 8.25% equal error rate in terms of correctly identifying each network with the clients satisfied with the results. This concept can be applied to a wide variety of fields including studies with brain networks and even social media data for relevant pandemic tracing. Next steps for the project should be to implement machine learning models to the pipeline like Graph Neural Networks to reduce the very intensive calculations and find ways to include social media data in the graph networks.

1 Introduction

When a person replaces their phone number, the only aspect that changes in the person’s life is the number itself as their contact numbers and their social circles will most likely remain consistent. With the same social group kept around this person, there can be a noticeable number of repeated interactions between them through their phones. A certain pattern of life can be found and mapped out as the repetition of events continues, limited by the scope of the interactions done with the phone number. With every individual having a different pattern of life in some way or another, such patterns can be used as key identifiers regardless of the phone number in use.

This is one of many cases of re-identification. The main idea of this concept is to take anonymized yet unique data and identify the individuals who own it by matching the data with entries from fully identified databases. Such databases would include other instances similar to the anonymized data but with an identifiable label. With the scenario above, the pattern of life extracted from the interactions from the phone would be the anonymized data and the name of the owner for the phone will then be entries from a known database.

One method for storing such complex networks is to use the graph data type. Consisting of nodes representing various entities and edges setting relationships between those nodes, graphs are able to handle containing such non-linear data. There are attributes within the nodes themselves to add unique characteristics and weights to the different nodes, allowing for graphs to be unique.

Law enforcement often faces the problem of reidentification with phone numbers in their cases as the suspects would be swapping their phone numbers, resulting in the difficulty of identifying who is who when looking at call and messaging logs. Currently there is no working solution to this issue, so during my internship I was tasked to create a prototype data pipeline for the client that would extract an interaction-based graph network from call and message logs and compare between various networks to find the most probable match to a specific phone number.

2 Review of Research

The need for graph matching was vital for the solution of the project, resulting in research in that area. This leads to an important piece of work that is actually implemented in the pipeline. The NetLSD package, created by Tsitsulin and his team, is a Network Laplacian Spectral Descriptor that extracts unique descriptors based solely on the graph's structure and allows for straightforward comparisons of large graphs, outperforming in efficiency and expressiveness (Tsitsulin 2018). This powerful tool is used within the pipeline to provide a point of comparison for the large interaction networks generated from the phone numbers.

Peter and Francois provide an overview on the concept of graph matching and the various techniques using different distance measures (Wills 2019). Though the techniques in this source were not used for the actual calculations in the pipeline, it served as a good starting point as the authors go over topics that are found in other sources such as spectral distances and explain them in a simplified manner.

Another relevant work that provides more context to the data was published in 2015 by a team consisting of Blondel, Decuyper, and Gauiter. The three go over the analysis of social graph networks extracted from anonymized data accumulated over the past decade and explore the vast array of ways this data can be used. This includes urban sensing, tracking of epidemics and geographical partitioning (Blondel 2015). Much of the information found in the trio's work provided the possibility of where to take the project next as well as providing suggestions on how to format the graphs in the pipeline as well.

3 Project Design

To understand the data pipeline, it's important to understand the project's constraints and the technology used. The client required that the prototype for the pipeline be in Python3 and run as a script on a command line. Cloud services were not recommended due to the sensitivity of the data. The pipeline should be able to parse in CSV files provided. Also, the data used by the client were not labeled, so machine learning algorithms like Graph Neural Networks that require labeled data were not feasible. From the sample data used to test the pipeline, each CSV contains call and message logs of one phone number with each entry containing the date, source phone number and destination phone number. Python includes various data analysis packages that are used, including Pandas and Numpy. NetworkX is another Python package that contains the Graph data structures and graph distance algorithms that are used to create the networks. NetLSD, discussed in the Review of Research section, is also used alongside the distance algorithms for the calculations for determining graph matches. Matplotlib was also used for visualizations of the graphs.

The pipeline has 3 distinct stages: parsing in CSV files, extracting data into graph networks, and finally calculating the distances from other networks. In the first stage, the CSV files are formatted in a specific way that a simple Python could be used to parse in data into a Pandas Dataframe. The second stage runs a function developed to loop through the Dataframe and create a NetworkX graph with all the phone numbers as nodes and the interactions between the numbers as edges.

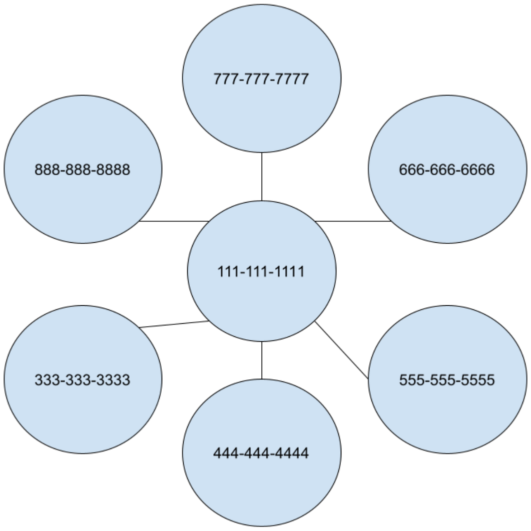

Figure 1: An Example of a Network

An example of an extracted graph network is shown in Figure 1 where the source phone number is 111-111-1111 and the adjacent nodes represent the numbers interacting with the source number. The final stage compares the newly created network to a list of precompiled graph networks, which will be referred to as PG, to determine the most probable match. This stage uses a combination of two distance measures, Graph Edit Distance from NetworkX and the NetLSD, using a simple sum. The Graph Edit Distance returns the number of changes required to convert from one graph to another where changes include adding or removing nodes, attributes, and edges. This stage runs both comparison algorithms to the new network with PG. The values are normalized and added together. This stage then returns a Dataframe with each entry containing both graphs compared and the sum of the normalized distances. The most likely to match will have the lowest distance sum score.

Challenges during this project include having no prior experience with the concept of Graph Matching, a fellow intern not well versed with Python, and having to work remotely due to Covid-19. Reading papers and frequently asking questions of the managers was necessary in order to understand the concepts of graph matching, bringing us interns to the same page. Also, assisting other team members and myself with Python eventually led my fellow intern to be a sufficient coder near the end of the internship. The pandemic also hindered work schedules as the same intern was located across the country, and the rest of the team was out of state as well. However, the team members provided a welcoming online experience, resulting in overcoming the distance and time zones and growing as a unit.

4 Results

The metric used to measure the performance of the data pipeline was the equal error rate. This is the location on the Receiver Operator Characteristic Curve, a plot that shows the diagnostic ability of a binary classifier system, where the false acceptance rate is equal to the false rejection rate, with a lower score resulting in higher accuracy. The pipeline had an 8.25% equal error rate which means it has a good accuracy score that could most definitely be optimized with further adjustments. Stage 3 of the pipeline takes most of the run-time, mainly due to NetworkX's Graph Edit Distance. This algorithm is a NP-Hard problem that can't be optimized by a significant amount. Thus the pipeline requires 10-15 minutes to complete its output. However, with a fairly low equal error rate, the client was satisfied with the outcome and pushed for further research into this project.

5 Conclusion

This data pipeline was created with a simple Python script to help a law enforcement client with reidentifying unknown phone numbers based on patterns of life. It is important to remember that the purpose of this pipeline is solely for law enforcement cases rather than commercial use so lawful citizens would not have to worry about their privacy. This technology could also be used for other industries such as pandemic tracking and brain neuron mapping ~~as well~~.

6 Future Work

Future work includes optimizing the combination of the two distance measures, finding ways to utilize Graph Neural Networks, and using more in-depth data such as social media data. The current method of combining the Graph Edit Distance and NetLSD is a simple sum, which could be optimized to a better combination. This can be done with adding weights adding more attributes to the nodes in the networks to help give more distinguished comparisons. In the long run, Graph Neural Networks can be implemented to replace the inefficient Graph Edit Distance algorithm with most of the work going to labeling the data missing from the beginning of the project. This pipeline can be modified and used to experiment with different types of data as well. Exploring rich data like social media can help improve the performance of the pipeline.

7 UVA Course Evaluation

Many of the higher-level courses at UVA promote large amounts of teamwork and project experience that had been applied to this project. Such classes include CS 4750 (Database Systems), CS 3240 (Advanced Software Development), and CS 4640 (Web Programming Languages). Along with teamwork and time management skills, I gained experience with using Big Data and the technology with it in CS 4774 (Machine Learning) that prepared me for a main part of this project.

However, none of my coursework really prepared me to analyze research papers effectively and this took too much time to read during my internship. Providing more hands-on experience with research papers would probably be the one change I would suggest to the UVA CS curriculum.

STS Topic

Facial Recognition Technology (FRT) has become an essential part of everyday life as it is being used on a daily basis with the norm of being used to unlock smartphones and other modern devices. China, however, is taking it a step further in the extreme of how they are implementing FRT, both commercially and governmentally, at the cost of personal information and security. With little to no presence of regulation of FRT, companies and the Chinese government had no limitations to what they can add FRT to, resulting in the information-sensitive technology being found in various processes such as making payments all the way to accessing parks and other public areas. Utilizing the Actor-network theory found in Redfield, this paper will be looking at the rise of FRT in China and find out how the ecosystem consisting of the citizens, government, and industry all interweave and allow for FRT to thrive. This dystopian-like scenario is an important case study as it can provide insight to other countries around the globe on the effects of heavily using peoples' privacy as it has on the ecosystem.

FRT's main role is to authenticate the identity using biometric data extracted from an individual's face to provide better security and convenience. This task is done by utilizing automated face detection and analysis software that takes into account various features of the face, including the eyes, nose, and mouth. These personal identifiers are then extracted into manageable data and compared to a database to finally match and authenticate whatever needs to be processed (Zhang 2021). The data is being extracted from the users, or in this scenario the citizens, by the industry and government. Both of them however are using it for different motives. To summarize, there are three actors at play here: the citizens who are actively using FRT in their daily lives, the authoritative government who utilizes FRT to monitor the citizens and maintain security and control, and the privatized industry who provide the citizens various implementations of FRT to use.

With the lack of regulation for FRT from the government and the advancements that occurred with FRT in 2014, China saw various tech companies and startups rising up and taking advantage of this sudden gold-mine of an industry (Lee 2021). The industry soon after pushed out more than a handful of FRT tools to citizens to the point where actions ranging from making payments to accessing parks and facilities are handled with FRT (Zhang 2021). Though these companies are handling such sensitive personal information from citizens, they continually add more integrations with FRT with their services and products.

A major reason why FRT from industry had taken over China's culture was due to people accepting the new technology and entrusting such personal data for two key factors; improved security and more convenience (Brown 2021). This culture shift and lack of regulation only grew the industry's presence and began to incorporate FRT in ways that lacked convenience and infringed on privacy. This is seen with China's first lawsuit case where a law professor fights against a company for registering him for facial recognition without consent. The ruling from the Government with its judicial court, however, did not cover what grounds private industries can utilize FRTs but did indicate what circumstances industries should obtain consent before collecting sensitive facial data (Brown 2021). Tension between the industry and citizen actors is not foreign in this network.

A study with 4 countries including China was conducted with online surveys to gauge the acceptance of facial recognition usage by institutions. From the 6100 Chinese citizens surveyed, the acceptance for FRT use by private enterprises is 17%, private-public partnerships (PPP) is 58%, and central government is 60% (Kostka 2021). There is a large difference in percentage between the citizen's acceptance in government and private industries, revealing that the authoritarian government has more control and moderation in their use of FRTs. China's government mainly utilizes the FRT for surveillance to monitor and track to increase security or provide convenience in public areas like automated traffic in metros. This appeals to the citizens as the government frame itself to serve the people, vastly different from the pro-profit industry. It is important to state the authoritative nature of the Chinese government as it could have affected the answering of some of the participants. This explains why many public debates about FRT are about private industry usages rather than public institutions.

With further problems arising due to private industry's personal data usage with FRT, the government is in talks of large regulations of FRT while taking further steps to continue furthering their image to the public as an ally (Zeng 2019). Though there is some criticism found in how the government is handling certain FRT scenarios including integration in schools and public transportation, such disputes are rather about the method of implementation of the technology with the critics framing themselves to be aligned with the state. As seen with the first lawsuit against FRT mentioned above, it is shown that the government wants to remain the ethical actor when compared to the industry as they only add limited regulations that weakly protect the citizens (Brown 2021). Used as a shield, the industry would then take the flak from the public as they dig more into the industry's "extreme" usage of FRT, building more trust with the Government with their "neutral" stance with FRT.

With the increasing pressure from the government and citizen actors, the country is seeing the industry actors beginning to make changes and making regulations of their own. Top FRT companies including Megvii and Alipay are creating initiatives and committees to support more normative and controllable facial recognition data safer for the citizen's privacy (Zeng 2019). This begins to show the cracks of this network ecosystem with the three actors. As the industry becomes tamer with regulations, it begs the question when the industry becomes as grounded as the government will the people start to shift their focus on disputes to the government.

This research is important due to the fact that FRT is being used all around the world and adapting at a tremendous rate and utilizing this unique case where usage of FRT as an extreme to show other countries how such an ecosystem between an authoritative government, private industry and citizens are handling the mass use of FRTs and whether it is viable in such situations.

Next Steps

As mentioned in the individual research section, the development of the pipeline was completed during a prior internship and is up to the client's wishes for further development in the technology. The capstone course as well will wrap up by the end of the semester with the final edit being turned in on November 7th, 2021. Following this prospectus, further research in the STS topic will be done in the 4600 course in the spring semester. For the STS section, there are many branches I would like to explore. A dataset that was used in a source had only taken ages from 18-27, so comparing how that age in other countries react to FRT would be next. Along with that, exploring how other countries themselves utilize and regulate FRT would bring more scope and provide more context to the idea prior.

References

Blondel, V. D., Decuyper, A., & Krings, G. (2015, February 11). *A survey of results on mobile phone datasets analysis*. arXiv.org. Retrieved October 28, 2021, from https://arxiv.org/abs/1502.03406.

Brown, T. G., Statman, A., & Sui, C. (2021). Public Debate on Facial Recognition Technologies in China. In MIT Case Studies in Social and Ethical Responsibilities of Computing. PubPub. https://doi.org/10.21428/2c646de5.37712c5c

Dou, E. (2021, July 30). China built the world's largest facial recognition system. now, it's getting camera-shy. The Washington Post. Retrieved October 4, 2021, from https://www.washingtonpost.com/world/facial-recognition-china-tech-data/2021/07/30/404c2e96-f049-11eb-81b2-9b7061a582d8_story.html.

Fadillah, D., Nuryana, Z., & S. (2020, April 25). Public Opinion of the facial recognition policy in China by Indonesian Student in Nanjing City. https://doi.org/10.37200/IJPR/V24I4/PR201272

Kostka, G., Steinacker, L., & Meckel, M. (2021). Between security and convenience: Facial recognition technology in the eyes of citizens in China, Germany, the United Kingdom, and the United States. Public Understanding of Science, 30(6), 671–690. https://doi.org/10.1177/09636625211001555

Lee, S. (2021, September 30). Coming into focus: China's facial recognition regulations. Coming into Focus: China's Facial Recognition Regulations | Center for Strategic and International Studies. Retrieved October 4, 2021, from https://www.csis.org/blogs/trustee-china-hand/coming-focus-chinas-facial-recognition-regulations

Tsitsulin, A., Mottin, D., Karras, P., Bronstein, A., & Müller, E. (2018). NetLSD. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3219819.3219991

Wills, P., & Meyer, F. G. (2019, December 17). *Metrics for graph comparison: A practitioner's guide*. arXiv.org. Retrieved October 28, 2021, from https://arxiv.org/abs/1904.07414#.

Zeng, Y., Lu, E., Sun, Y., & Tian, R. (2019, September 19). *Responsible facial recognition and beyond*. arXiv.org. Retrieved October 14, 2021, from https://arxiv.org/abs/1909.12935.

Zhang, L.-L., Xu, J., Jeong, D., Ekouka, T., & Kim, H.-K. (2021). The effects of facial recognition payment systems on intention to use in China. *Journal of Advanced Researches and Reports*, *1*(1), 33–40. https://doi.org/10.21742/jarr.2021.1.1.05