Detecting Criminal Networks via Non-Content Communication Data Analysis Techniques from the TRACY Project

Pradeep Rangappa[1], Amanda Muscat[1,4], Alejandra Sanchez Lara[1], Petr Motlicek[1,3], Michaela Antonopoulou[2], Ioannis Fourfouris[2], Antonios Skarlatos[2], Nikos Avgerinos[2], Manolis Tsangaris[2], and Kasia Kostka[5]

[1]Idiap Research Institute, Martigny, Switzerland
{pradeep.rangappa, amanda.muscat, alejandra.sanchezlara, petr.motlicek}@idiap.ch
[2]Performance Technologies, Anonymos Etairia Pliroforikis, Athens, Greece
{Michaela.Antonopoulou, Giannis.Fourfouris, Antonis.Skarlatos, Nikos.Avgerinos, Manolis.Tsangaris}@performance.gr
[3]Brno University of Technology, Czech Republic
[4]University of Malta, Msida, Malta
[5]Timelex, 1000 Brussels, Belgium

Abstract. This paper explores the critical role of non-content data (NCD), provided by electronic communications service providers in aiding criminal investigations. As highlighted by the Law Enforcement Agencies (LEAs) and the European Commission, NCD plays a fundamental role in identifying suspects and discerning behavioral patterns. Despite its significance, LEAs encounter various challenges in effectively analyzing the extensive volume of NCD. To address this issue, this paper presents the importance of (although simulated but realistic) data collection, the technologies that can be built and the methods for detecting the suspect within the framework of the TRACY project. These techniques aim to enhance capabilities of LEAs by processing large-scale NCD and aligning it with existing evidence. By prioritizing the tracing of suspects movements and integrating data from diverse NCD sources, TRACY's initial approach on synthetic data promises to significantly advance the identification of offenders involved in serious and organized crime.

Keywords: TRACY · Law Enforcement Agencies · Suspect Detection· Non-Content Data· Social Influence Analysis· Link Prediction

1 Introduction

Lawful access to communication data stands as a cornerstone during a criminal investigation and prosecution. Reports from Law Enforcement Agencies (LEAs) [1] have stated that Non-Content Data (NCD), or metadata, provided by Electronic Communications Service Providers (CSPs) play a vital role into an investigation or prosecution of organised crime and terrorism cases. Statistical analyses indicate that non-content data have been requisitioned in no less than 60% of cases over the preceding two years across all European Union (EU) Member States [1]. However, analysing such data remains a complex task, owning to its voluminous nature and the lack of knowledge that is present amongst LEAs in the domain of data processing through artificial intelligence modelling. This knowledge gaps hinders on LEAs' efficacy in harnessing non-content data to their advantage. In successful cases, LEAs collect non-content data and attempts in building a profile of the suspect by establishing the general behavioural patterns from geographical locations and/or build a network to assess the most probable locations of the suspect's presence. This work is part of the TRACY (A big-data analyTics from base-stations Registrations And Cdrs e-evidence sYstem) project[1], a European Union research initiative furnishing LEAs with a digital forensic and investigative platform tailored to identifying suspects with reference to temporal and spatial parameters.

Non-content communication data, essentially linked to a subscriber or customer, represents a key offering provided by ESPs during the course of criminal investigations. According to the 2020 report issued by the European Commission's Directorate-General for Migration and Home Affairs [1], non-content communication data can be categorized into three primary categories:

- Subscriber data: comprising information enabling the identification of the sender of a communication (e.g., name, address, phone number).
- Traffic data: encompassing details requisite for identifying the type, date, time, and duration of a communication, along with any information facilitating the identification of the receiver(s) or attempted receiver(s) of said communication.
- Location data: encompassing details essential for pinpointing the location of the communication equipment (e.g., cell tower location).

This type of data is generally retained by CSPs across most EU Member States for at least one internal purpose, such as marketing and invoicing. The mandated data retention period for law enforcement purposes stands at 12 months, with exceptions in Ireland and Italy [1]. Furthermore, beyond law enforcement, non-content data finds application in the private sector, aiding companies in providing personalized services and facilitating targeted advertising campaigns on social networking platforms [2].

Recent studies have shown the importance of analyzing mobile phone data for criminal investigations. For example, [3] conducted a systematic review highlighting how mobile phone data can be used to study human communication behaviors and mobility patterns to detect criminal networks and predict crime. Additionally, research by [4] has underscored the effectiveness of network-based approaches in enhancing the predictive accuracy of crime detection algorithms. Moreover, [5] explored how NCD, when analyzed through advanced data processing techniques, can significantly improve the identification of criminal activities and suspects. Non-content communication data is typically analysed using a

[1] https://www.tracy-project.eu

structure resembling a graph consisting of nodes representing individuals, devices (i.e., mobile phones), or cell IDs and edges denoting connections between said entities. Networks are used for a multitude of purposes, including link prediction [6, 7] and social influence analysis [8].

This paper proposes a new set of techniques built on processing large amounts of non-content data aligned with existing evidence available from other means. The data processing techniques are expected to offer an innovative digital solution to advance the process of LEAs investigations and prosecutions in determining the general behavioural patterns of the offender, assisting in the identification of the offender in serious and organised crime when dealing with non-content data provided by ESPs. The technologies are being developed as part of TRACY project, further expanded on existing knowledge gained in recently ended EU funded collaborative research and innovation project, ROXANNE[2] (Real time network, text, and speaker analytics for combating organized crime). The aim of ROXANNE was to unveil criminal networks from call content data through the integration of speech and language technologies [9] [10].

More specifically, TRACY's technology is built on processing large amounts of non-content data available from the area correlated with a crime for a given time period. Relatively simple but computationally-heavy heuristics are investigated as part of TRACY project to trace movement of mobile devices connected to the telecommunication network and align them with existing evidence available from other sources (e.g., CCTV footage captured near the crime scene area, covering the incident by unidentified individuals, potentially indicating the direction of their movement taken after the crime).

The paper is organised as follows: Section 2 explains the dataset utilized to simulate non-content communication data derived from operational cases, the TRACY dataset. Following this, section 3 provides a realistic scenario where such analysis can be used. Section 4 introduces the technologies. Subsequently, section 5 presents the outcomes of our experiments. Finally, section 7 encapsulates our concluding remarks.

The motivation driving this study is to sift through the extensive data sources available during an investigation and establish a method to prioritize relevant information. This mirrors the challenges encountered in real crime investigations, where LEAs are faced with vast amounts of data which can quickly become overwhelming for them and potentially loosing the focus on important evidence. The scenario described hereunder gives a clear and realistic example of how the technologies proposed in this paper can be utilised in real-life cases

2 Dataset

This section introduces the TRACY dataset, formed by merging data from SIMCITY and SIMNET simulators, for crime scene analysis.

[2] https://www.roxanne-euproject.org/

The SIMCITY simulator specializes in generating synthetic data that mimics user movement behavior. To achieve this, it combines data generated from the integration of three distinct components:

1. The Multi-Agent Transportation Simulation (MATSim) [11] is utilized, to replicate drivers behavior. MATSim is tailored for modeling complex transportation systems and individual agent behaviors within such networks. The simulator relies on two primary inputs: the transportation network and the activity plans. The transportation network, sourced from OpenStreetMap [12], details road characteristics such as capacity and speed limits. An activity plan outlines the sequence of activities an agent will perform throughout a day, including where and when these activities will occur. Examples of activities include going to work or school, taking a stroll and so on. It is worth noting that the proportion of people engaging in each activity is derived from publicly available data from organizations like Hellenic Statistical Authority [13]. The actual route taken by an agent is computed dynamically by MATSIM, considering factors like road usage, congestion and travel times.
2. A custom pedestrian simulator, accurately emulates streetwalker movements, addressing the limitations of MATSim in simulating pedestrians due to the incomplete sidewalk information in Athens' transportation network. This simulator, implemented in Python, utilizes the OSRM library [14] to compute walkers' movement based on a plan similar to the aforementioned activity plan. Each individual is simulated to take up to three walks at different times, throughout the day.
3. A Python Criminal Ingestor Simulator utilizes again the OSRM library to generate routes for simulated criminals based on predefined location pairs and timestamps, representing pieces of evidence like crime scenes or CCTV footage of the criminals. The generated routes serve as the ground-truth and are used to assess the effectiveness of the implemented algorithms.

Overall, the SIMCITY simulator generates GPS trajectories and travel durations, which are crucial for examining person movements. Given our focus on information related to a crime scene and the need to reduce the size of the stored data, we save only the trajectory data of individuals who meet certain criteria: they must have moved within a designated timeframe corresponding to the approximated time of the crime and passed at least once through a bounding box that defines the surrounding area of the crime scene.

SIMCITY was executed on one million individuals. Based on a timeframe of 08:00 to 12:00 and a predefined bounding box in the Kerameikos, Athens, Greece (Fig. 1). Subsequently a dataset comprising GPS trajectories and associated timestamps of 30,703 individuals was obtained. Furthermore, the SIMNET simulator was utilized to augment the data with network cell service information, improving dataset authenticity with realistic cellular communication patterns corresponding to individuals' movements. The simulator, implemented in Python, works on top of the SIMCITY dataset. Its core functionality involves assigning a serving cell to each user's GPS point, depending on the Received

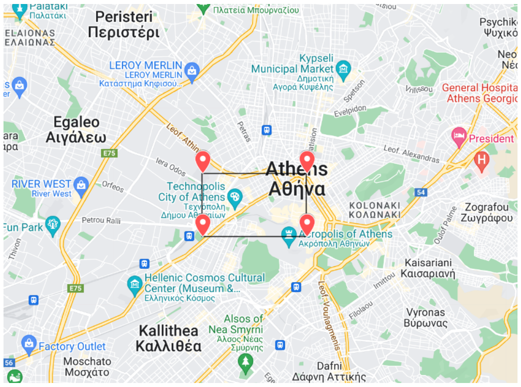

Fig. 1. The ‘Kerameikos’ bounding Box.

Signal Strength Indicator (RSSI) which is calculated from factors like the cell’s distance, orientation etc. and initiating a handover with the target base station when the calculated RSSI weakens below a predefined threshold. This approach aims to reflect the real-world behavior of cellular networks by capturing the complexities of mobile telecommunications.

SIMNET employs network topology data from open sources datasets like OpenCellid [15] and Unwired Labs [16], supplemented by reasonable assumptions regarding antenna direction[3], coverage and tilt, which are used in the RSSI calculation. A visual representation of the final generated data can be seen in Fig. 2, where we showcase the movement of a specific person. The blue points

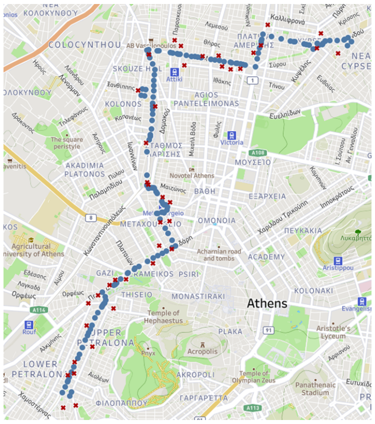

Fig. 2. Trajectory snapshot showing the movement and the serving cells of a specific person.

[3] https://www.cellmapper.net/

represent the actual GPS coordinates of the person's movement, while the red points showcase the serving cells.

Table 1 provides a comprehensive description of the synthetic dataset.

Table 1. Description of the synthetic dataset.

Column Name	Description
person_id	The Individual's unique identifier
lat	The Latitude of the individual at the recorded timestamp
lon	The Longitude of the individual at the recorded timestamp
measured_at	The Timestamp of the measurement
cellid	The Serving Cell ID for the specific latitude and longitude point
cell_lat	The Latitude of the servicing cell
cell_lon	The Longitude of the servicing cell

The dataset includes $'person\_id'$, a unique identifier for each individual, along with their geographical coordinates, $'lat'$ and $'lon'$, which specify their exact location at the recorded time, $'measured\_at'$. Additionally, the dataset contains the $'cellid'$, identifying the serving cell for each location point and the coordinates $'cell\_lat'$ and $'cell\_lon'$ that pinpoint the location of the serving cell.

2.1 Privacy and Ethics in Data Collection and Processing

Analysis of the non-content data comes with several ethical implications, particularly when it comes to privacy and data protection. Although non-content data does not involve the actual content of communications, it can reveal sensitive information about the data subject's identity or location and therefore, can be misused. Consequently, while it can be of immense value to LEAs and help in identifying suspects and discerning behavioural patterns, rights such as respect for private and family life, protection of personal data and freedom of expression and information need to be respected at the same time.

In order to do so, a principle of proportionality should be taken into consideration while analysing non-content data [17]. However, it is often not an easy task to balance the right of individuals to the protection of their personal data while aiding law enforcement authorities in preventing, investigating, detecting or prosecuting criminal offences as well as guaranteeing a high level of public security. Additionally, transparency and accountability surrounding the methods of data analysis are an important aspect of upholding public trust and understanding. Nevertheless, the extend of such a transparency should not simultaneously jeopardise ongoing investigations. Regular training and audits are also an important aspect in preventing potential biases and prejudices and thus averting distrust and social tensions.

3 Scenario: A realistic application

In this section, we present a real-like scenario demonstrating the application of AI technologies utilizing the dataset outlined in the previous section.

Suppose a criminal activity occurs at point X and footage from video surveillance reveals two unidentified individuals, referred to as the suspects, fleeing the crime scene in the direction of point Y. Relying solely on CCTV footage proves to be a dead-end, thus alternative investigative strategies become imperative. Our objective is to leverage the extensive dataset of NCD to reconstruct the route taken by these unknown individuals by processing NCD and eventually filtering out individuals who did not travel in the same direction as the suspects. Additionally, we aim to explore the relationships between individuals based on their frequency of encounters (as gathered by NCD) within a four-hour timeframe. By employing the heuristics (to pre-process the non-content data), we generate network of individuals to be subsequently analysed by network-analysis technologies outlined below. Our goal is to uncover critical insights that aid in the identification and apprehension of the suspects, as well as elucidate social connections within the dataset.

4 Technologies

This section presents two different technologies that are transferred from ROXANNE project and can be adapted to TRACY project.

4.1 Social Influence Analysis

The first network analysis algorithm considered in this paper offers an innovative digital solution to advance the process of LEAs investigations and prosecutions within the TRACY project known as *Social Influence Analysis*. Social influence approaches are effective at encouraging resource conservation when compared to both a control group and alternative interventions [18].

The default setting for social influence analysis is the Pagerank algorithm. This algorithm is used to determine the relevance or importance of the dataset. It introduces the influence of a set of data present in a network to the computation of the centrality of the nodes [19]. The social influence analysis measure the importance of each entity based on the topology of the social network. It is possible to obtain a 'Social Influence Score' represented as a positive numerical value, indicating the degree of influence exerted by a node. Higher values signify greater influence of the node in question. TRACY project proposes to adapt this technology considering the identification, definition and creation of a network with respect to the synthetic data generated.

4.2 Link Prediction

The second network analysis algorithm considered in TRACY project is *link prediction*. Link prediction has been one of the most active research domains in network science including local similarity indices, link predictability, network embedding, matrix completion, ensemble learning and some others [20].

The default setting for link prediction is the Jaccard Coefficient algorithm. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets [21]. This algorithm is considered in this paper to assess similarities within the dataset for the purpose of conducting and predicting links referred to as $('perform\_link\_prediction')$. The module used as an entry point is the dataset. The parameters required for link prediction in this project are $'community\_detection\_method'$ and $'modularity'$. Once a perform link is created, the technology runs link prediction module and the results are saved as a JSON file. This file includes information such as the type, ID and different properties of the predicted links.

Link prediction is a feature of *Network Analysis*, and is available during the case analysis phase (after data file processing is complete on the platform). This technology is composed by community detection, social influence and outlier detection. The user can configure (able or disable) the technologies related to Network Analysis in Roxanne platform.

The figure depicted above (Fig. 3) illustrates how the application of link prediction techniques can be used on the TRACY dataset. In this representation, each node corresponds to a unique person_ID, while the edges between them signify the frequency of their encounters within a specified time-frame. The thickness of these edges reflects the number of meetings between two individuals, with thicker edges indicating more frequent interactions. Consequently, the frequency of meetings between two person_IDs can potentially serve as an indicator of the strength of their relationship. This network can then be used to

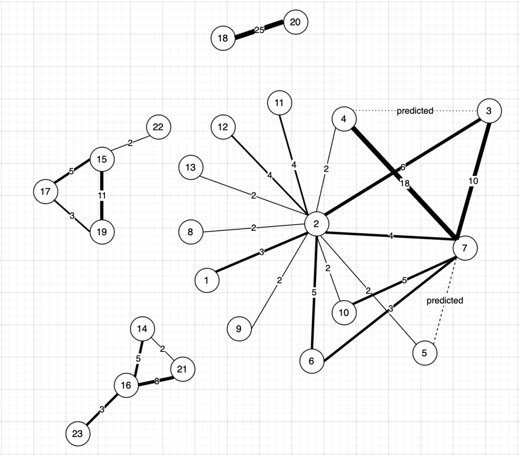

Fig. 3. Illustration of applying link prediction to a network of individuals.

predict whether two individuals, with no documented evidence of interaction in the dataset, could possibly be acquainted. Link prediction algorithm leverages the network's topology to predict precisely this. The output of the link prediction process would be denoted by dotted lines as shown on the plot in Figure 3.

5 Experiments

In this section, a systematic approach is proposed to find the consistent pair of individuals across various locations and times using synthetic data. The detailed process involves reading the syntheic data, segmenting it into time slots, identifying unique GPS locations, determining the number of individuals at each location, forming pairs, and then storing and updating the results. The following sub-sections highlights the proposed methodology to identify the suspects and filtering them effectively.

5.1 Exploring Consistent Pair Meetings Across Locations

Here we examine the consistency of pairs gathering across various locations as shown in Figure 4.

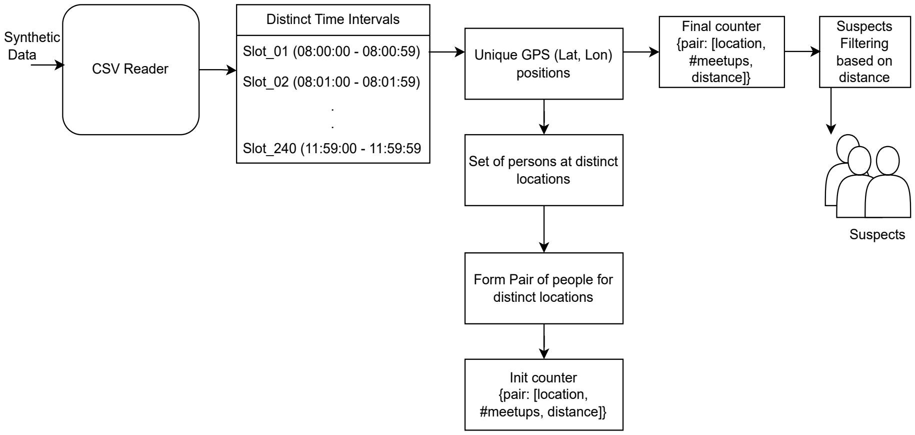
Synthetic Data
CSV Reader
Distinct Time Intervals
Slot_01 (08:00:00 - 08:00:59)
Slot_02 (08:01:00 - 08:01:59)
Slot_240 (11:59:00 - 11:59:59
Unique GPS (Lat, Lon) positions
Set of persons at distinct locations
Form Pair of people for distinct locations
Init counter {pair: [location, #meetups, distance]}
Final counter {pair: [location, #meetups, distance]}
Suspects Filtering based on distance
Suspects

Fig. 4. Functional Blocks to Detect Suspects from the Synthetic Data

1. **Reading Synthetic Data:** The TRACY synthetic dataset, which includes columns for person IDs, latitude, longitude, and timestamps (formatted as hh:mm), is inputted to a CSV reader. This typically results in the formation of a pandas dataframe (referred to as df).

2. **Segmentation into Time Intervals:** The entire dataframe is divided into discrete time intervals, each lasting 1 minute. For example, the time interval slot-01 would contain all rows from the synthetic data that have timestamps between 08:00:00 and 08:00:59. This segmentation results in the creation of 240 slots, spanning the time from 08:00:00 to 12:00:00.
3. **Identification of Potential Suspects:** Within each 1-minute slot, the following four critical steps are carried out to identify potential suspects:
   (a) *Identification of Unique GPS Coordinates:* Within each slot, we identify unique GPS locations. For instance, unique coordinates like (37.971077, 23.418176'), (37.971367, 23.415569'), etc. represent different GPS locations where individuals have been detected.
   (b) *Counting Individuals at Each Location:* For each unique GPS location, we determine the number of individuals present at that spot during the given time slot. For example, the record (37.9710775991439, 23.4181762259198) » (A, B, C) indicates that individuals with IDs A, B, and C were gathered at that specific GPS location.
   (c) *Grouping Individuals into Pairs:* The unique person IDs identified at each location are grouped into pairs. This involves generating all possible combinations of pairs from the set of individuals at each GPS location. For instance, if three individuals (A, B, and C) are present, the resulting pairs would be [(A', B'), (A', C'), (B', C')]. In a scenario where five individuals are present, a total of 10 pairs would be generated, representing all possible combinations of two individuals from the group of five.
   (d) *Storing Results in a Dictionary:* The results from the previous steps are stored in a dictionary. Each key in the dictionary represents a unique pair of individuals, while the associated value contains information about the locations where the pair met, the frequency of their meetings at different locations, and the distance they traveled together. The great-circle distance between two GPS points is calculated using the standard haversine formula [22], which computes the shortest distance over the earth's surface.
4. **Iterative Process for All Time Slots:** The steps outlined in Step 3 are repeated for all 240 slots. As identical pairs are observed across various locations and times (with a 1-minute buffer), the dictionary is continuously updated. This results in a comprehensive counter that includes a list of unique pairs who have gathered at different locations at different times.
5. **Sorting the Dictionary:** Finally, the dictionary is sorted based on the total distance traveled by the pairs between 08:00:00 and 12:00:00. This sorted list helps identify the pairs who have covered the greatest distances together, which can be critical in understanding their movement patterns and potential interactions.

By following this structured approach, we can systematically analyze the patterns of gatherings and movements of individuals, providing valuable insights into their behavior and interactions across various locations and times.

5.2 Suspects filtering based on distance

Here, we categorize pairs according to the distances they've traveled. To achieve this, we begin by selecting a pair and compiling a list of unique locations visited by them. Subsequently, we calculate the haversine distance between each of these locations. Finally, we aggregate these distances to determine the total distance covered by the pairs within the time span from 08:00:00 to 12:00:00.

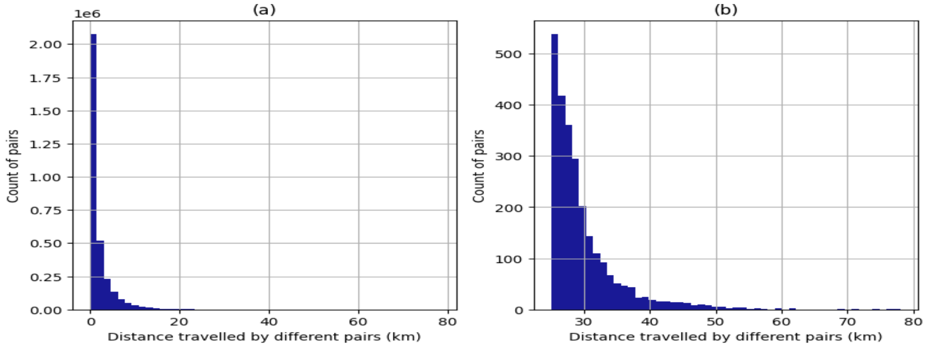

Fig. 5. (a) Histogram displaying the distances covered (in kilometers) by various pairs (having met at least twice).(b) Histogram illustrating distances traveled (in kilometers) exceeding 25 kilometers by different pairs (meeting at least twice).

Figure 5 (a) shows the histogram displaying the distances covered (in kilometers) by various pairs (having met at least twice), (b) shows the histogram illustrating distances traveled (in kilometers) exceeding 25 kilometers by different pairs (meeting at least twice). It can be observed that approximately 3 million pairs have traveled a distance ranging from 0 to 5 kilometers. as shown in Figure 5 (a). Likewise, there are 2564 pairs that have traveled distances exceeding 25 kilometers as shown in Figure 5 (b).

5.3 Visualizing the Suspects

In Figure 6 the plot was created by a large set of person ID's from the TRACY dataset. It can be appreciated how the interconnection of person ID's are connected to one another forming different networks. The plot was created using the synthetic TRACY dataset on Streamlit[4] as a graph[5] for an interactive approach. Streamlit is an open source platform that transforms Python scripts into interactive web apps, build dashboards, generate reports, and create chat apps. Agraph is based on react-graph-vis, a rendered graphs that are scrollable, zoomable,

[4] https://streamlit.io/

[5] https://github.com/ChrisDelClea/streamlit-agraph

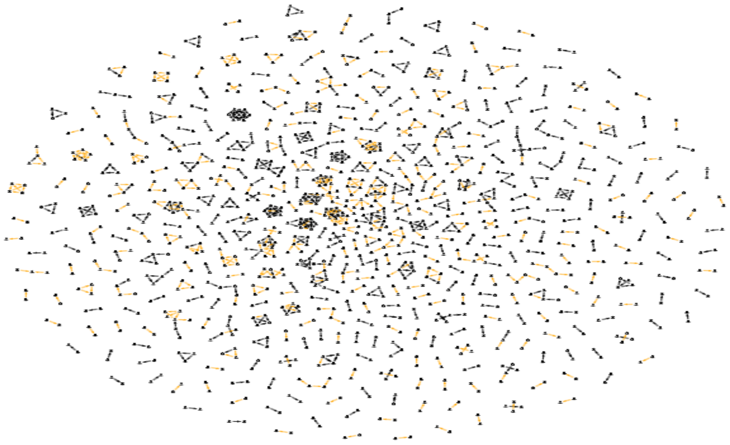

Fig. 6. A network diagram of the TRACY dataset.

retina ready, dynamic, and can switch the layout with a double click. This network diagram displays how TRACY dataset is distributed and how individuals behave.

In analyzing the network graph derived from the TRACY dataset, our focus centers on two individuals identified as suspects, labeled as person ID 961138 and person ID 722362 as shown in Figure 7. Notably, from a closer look at the network graph, one can observe that these individuals encountered not only each other but also additional individuals including person ID 321286 and person ID 625851, as illustrated in Figure 7. In order to explain the trajectories of these suspects, we utilize Google My Maps for visualization. Figure 8 illustrates the presence of both suspects, person ID 961138 highlighted in orange and person ID 722362 in yellow, at the crime scene location (marked as "1"). Furthermore, it illustrates the trajectories of both suspects following the incident, leading to their reunion at a final destination (marked as "2"). Additionally, the presence of other individuals at the crime scene, person ID 321286 and person ID 625851, is indicated by blue and green markers, respectively.

6 Conclusions

In this paper, we examine the importance of non-content data offered by electronic communications service providers in assisting criminal investigations. It outlines the methods employed to generate the simulated dataset and demonstrates a practical application of different technologies such as social influence analysis and link prediction on this dataset. Additionally, the methodology for obtaining the consistent pair meetings across diverse locations is discussed. These approaches offer law enforcement agencies a valuable tool to filter and discard

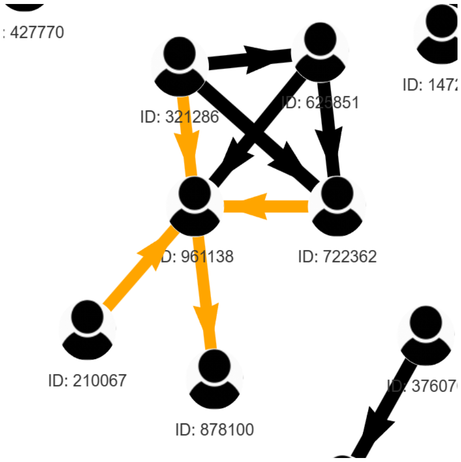

Fig. 7. A close-up snapshot of the network of person IDs identified as suspects in the experiments conducted.

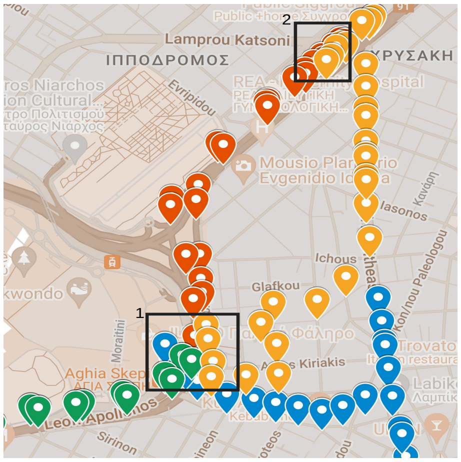

Fig. 8. A snapshot of Google My Maps capturing the crime scene location (marked as "1"), the trajectories taken by both suspects post-incident, and their reunion at a final destination (marked as "2"), Additionally, blue and green markers denote other individuals present at the crime scene.

irrelevant person IDs from the data and enable investigators to focus on the persons of interest.

Acknowledgements This work was supported by the European Health and Digital Executive Agency (HADEA) under the Commission Digital Europe Programme (DIGITAL) with grant agreement No. 101102641 (project TRACY - a big-data analyTics from base-stations Registrations And Cdrs e-evidence sYstem, 2023-2025).

References

1. Claire Dupont, Valentina Cilli, Ela Omersa, Camille Borrett, Maxime Moulac, Plixavra Vogiatzoglou, and Svetla Nikova. Study on the retention of electronic communications non-content data for law enforcement purposes. *Study on the retention of electronic communications non-content data for law enforcement purposes*, 2020.
2. Daniel Shickich. What your tweet doesn't say: Twitter, non-content data, and the stored communications act. *Wash. JL Tech. & Arts*, 8:457, 2012.
3. Cang-Hong Jin, Dong-Kai Chen, Fan-Wei Zhu, and Ming-Hui Wu. Detecting suspects by large-scale trajectory patterns in the city. *Mobile Information Systems*, 2019(1):1837594, 2019.

4. Sandrine R Müller, Joseph B Bayer, Morgan Quinn Ross, Jerry Mount, Clemens Stachl, Gabriella M Harari, Yung-Ju Chang, and Huyen TK Le. Analyzing gps data for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science*, 5(2):25152459221082680, 2022.
5. Héctor Cogollos-Adrián, Santiago Porras-Alfonso, and Bruno Baruque-Zanón. Software tool for analysis and visualization of gps tracks in urban environments. *Transportation Research Procedia*, 58:401–407, 2021.
6. Zahra Ahmadi, Hoang H Nguyen, Zijian Zhang, Dmytro Bozhkov, Daniel Kudenko, Maria Jofre, Francesco Calderoni, Noa Cohen, and Yosef Solewicz. Inductive and transductive link prediction for criminal network analysis. *Journal of Computational Science*, 72:102063, 2023.
7. Lucia Cavallaro, Annamaria Ficara, Pasquale De Meo, Giacomo Fiumara, Salvatore Catanese, Ovidiu Bagdasar, Wei Song, and Antonio Liotta. Disrupting resilient criminal networks through data analysis: The case of sicilian mafia. *Plos one*, 15(8):e0236476, 2020.
8. Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816, 2009.
9. Maël Fabien, Shantipriya Parida, Petr Motlicek, Dawei Zhu, Aravind Krishnan, and Hoang H Nguyen. Roxanne research platform: Automate criminal investigations. In *Interspeech*, pages 962–964, 2021.
10. Petr Motlicek, Erinc Dikici, Srikanth Madikeri, Pradeep Rangappa, Miroslav Jánošík, Gerhard Backfried, Dorothea Thomas-Aniola, Maximilian Schürz, Johan Rohdin, Petr Schwarz, Marek Kováč, Květoslav Malý, Dominik Boboš, Mathias Leibiger, Costas Kalogiros, Andreas Alexopoulos, Daniel Kudenko, Zahra Ahmadi, Hoang H. Nguyen, Aravind Krishnan, Dawei Zhu, Dietrich Klakow, Maria Jofre, Francesco Calderoni, Denis Marraud, Nikolaos Koutras, Nikos Nikolau, Christiana Aposkiti, Panagiotis Douris, Konstantinos Gkountas, Eleni Sergidou, Wauter Bosma, Joshua Hughes, and Hellenic Police Team. Roxsd: The roxanne multimodal and simulated dataset for advancing criminal investigations. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 17–24, 2024.
11. Kay W Axhausen, Andreas Horni, and Kai Nagel. *The multi-agent transport simulation MATSim*. Ubiquity Press, 2016.
12. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org , 2017.
13. Hellenic Statistical Authority. https://www.statistics.gr/.
14. Dennis Luxen and Christian Vetter. Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516, New York, NY, USA, 2011. ACM.
15. OpenCelliD. https://www.opencellid.org/.
16. Unwired Labs. https://my.unwiredlabs.com/.
17. EU EU. Charter of fundamental rights of the european union. *The Review of International Affairs*, 63(1147):109–123, 2012.
18. Thomas J Zagenczyk, Kristin D Scott, Ray Gibney, Audrey J Murrell, and Jason Bennett Thatcher. Social influence and perceived organizational support: A social networks analysis. *Organizational Behavior and Human Decision Processes*, 111(2):127–138, 2010.
19. Leandro Tortosa, Jose F Vicent, and Gevorg Yeghikyan. An algorithm for ranking the nodes of multiplex networks with data based on the pagerank concept. *Applied Mathematics and Computation*, 392:125676, 2021.

20. Tao Zhou. Progresses and challenges in link prediction. *Iscience*, 24(11), 2021.
21. Mubbashir Ayub, Mustansar Ali Ghazanfar, Tasawer Khan, and Asjad Saleem. An effective model for jaccard coefficient to increase the performance of collaborative filtering. *Arabian Journal for Science and Engineering*, 45(12):9997–10017, 2020.
22. Nitin R Chopde and Mangesh Nichat. Landmark based shortest path detection by using a* and haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):298–302, 2013.