EE-558: A Network Tour of Data Science

Final Project:

"Exposing the True Terrorist Network"

Dataset: Terrorist Relations

Timur Lavrov, Ayberk Tarçın, Sinan Yılmaz, Aslı Yörüsün
School of Engineering, EPFL, Switzerland
January 18, 2019

I. ABSTRACT

In this project, we work with the Terrorist Relations dataset that is taken from the Profile in Terror (PIT) knowledge base. As terrorists arguably present the largest threat to today's society, it is of interest to analyze and reveal interesting properties of their network in order to guide anti-terrorist campaigns and help ensure public safety around the world. This dataset is particularly interesting as data relating to terrorists is scarce due to their secretive nature. The PIT dataset is constructed from information extracted from public news sources. Hence, this dataset is purely built on empirical evidences and it safe to assume that it is incomplete. In this project, we try to predict the unobserved terrorist relations and present an expansion of the given terrorist network. Switching from a standard to a relationship graph representation, we then perform signal interpolation in order to predict the labels of the new relationships. Subsequently, we performed analyses on our networks with the aim to identify the top priority terrorist for anti-terrorism campaigns to target in order to successfully weaken the network.

II. TERRORISTREL DATASET

A. Motivation for the Dataset

The motivation for us to choose **TerroristRel** [1] dataset is to examine how terrorists are related to each other and how this affects the spread of terrorist ideas and terrorist activities. Analyzing the network of personal connections is important to understand how the terrorist organizations operate. Thus, if we have an idea about who knows whom and what type of relationship they have, the extent of the terrorist activities can be understood. Moreover, if we learn more about different organizations and their connections, this can provide new ways to tackle terrorism around the globe.

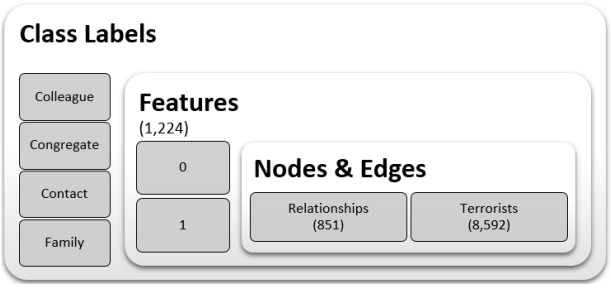


Fig. 1 Terrorists Relations Dataset

B. Data Exploration

The TerroristRel dataset contains a subset of the data containing terrorists and their relationships collected by the MIND Lab at UMD (http://www.mindswap.org/). The dataset contains 851 terrorist relationships (nodes), each described by a 0/1-valued vector of attributes where each entry indicates the absence/presence of a feature. In this dataset, the edges represent terrorists, and two nodes share an edge if the respective relationships have a terrorist in common. It was designed for classifying the type of relationship connecting two terrorists and contains two types of information. Also, the nodes are represented with non-existing

URL. We parsed these URLs in order to take the name of the terrorists. Some URL links does not have terrorist names, they only contain the time when they were extracted. Hence, some of the terrorists are represented as timestamps.

The TerroristRel dataset contains 4 different feature tables. Each of these tables assign the nodes one of the labels presented below:

- Colleague (54.2%): Two people are members of the same terrorist organization.
- Family (16%): Two people are in the same family (e.g. father-son, husband-wife, uncle-nephew, cousin-cousin).
- Contact (17.4%): Two people have contacted each other (e.g. attend the same meeting, email each other, call each other via phone).
- Congregate (12.4%): Two people use the same facility (e.g. went to the same training camp).

Additionally, the relationships are characterized by 1224 feature values (half of them [1-612] are for the first terrorist in the relation and other half [613-1224] for the second). Hence, we have 612 distinct attributes describing each terrorist in our network. Some of these binary valued features are extracted from keyword analysis of the terrorist biographies. [2]. A portion of the feature table is presented in Table I. However, the meanings of these 612 features were not specified in the dataset. TerroristRel dataset provided four different tables for every label mentioned above. In order to work on a single pandas dataframe, these 4 different tables were merged. Additionally, we did not perform down-sampling since data was already collected, organized and had small number of nodes.

C. Dual Graph

Throughout the semester, we analyzed our dataset by using the nodes and edges defined in original TerroristRel dataset. In addition to the original graph, we created a Dual Graph which includes individual terrorists as nodes and relationships as edges. The Dual Graph can provide more intuitive perspective for individuals and their respective relationships. We found out that there are 244 different terrorists in our original data. This gives us 244 nodes and 851 possible edges between those nodes in the Dual Graph. At this step, we ignored the labels so that our Dual Graph has only one edge between every two terrorists instead of two mirror-edges. We found out that there are only 11 multi-labelled relationships shown in Table II out of 851. Hence, Dual Graph will have 244 individual terrorists and 840 distinct relationships (edges) between those terrorists.

III. EXPERIMENTS

Since the origin of our data is based on various sources like news and media reports, only a portion of true terrorist network is represented in the original data. However, it is possible to discover a true terrorist network by using feature table. To understand this true network, we build potential relations between terrorists based on their respective 612 distinct features.

First, a feature table of Dual Graph is generated from merged feature table. This feature table contains 244 terrorists' feature vector. While forming the feature table, we realized that there are some irrelevant features that can be filtered out. Those filtered features include features for which all terrorists have binary value of zero and features that were highly correlated. One out of those highly correlated features are taken and rest is neglected. That's why we reduced the size of feature table from 244×612 to 244×440.

Subsequently, the correlation between terrorists' features are calculated to estimate hidden relationships. A high correlation coefficient was selected

Terrorist Relations		Features				
Terrorist ID_1	Terrorist ID_2	1	2	...	1223	1224
UmarBaziyani	SaifAldinMustafaNuaimi	0	0	...	0	0
2006/02/01/17:03:5	2006/02/14/16:53:51	0	0	...	1	1
2005/09/07/09:36:59	MustafaKamel	0	0	...	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

TABLE I Portion of Features Table

Terrorist Relations		Labels			
Terrorist ID_1	Terrorist ID_2	Colleague	Family	Congregate	Contact
Zawahiri	OBL	x			x
Moulad_Bouguelane	Lionel_Dumont	x			x
AhmadAjaj	EyadIsmoil	x			x
FathiKhatib	NasserWatimi	x			x
2006/02/02/16:11:52	2006/02/02/16:03:38	x	x		
FathiKhatib	MohammedSchreim	x			x
2006/03/17/05:47:04	2006/03/15/22:21:03	x			x
2005/12/22/19:57:48	Moulad_Bouguelane	x			x
MohammedSchreim	MuammarSheikh	x			x
FarouqQaddoumi	Yasser_Arafat	x			x
MuammarSheikh	NasserWatimi	x			x

TABLE II Terrorist Relations that have multi-labels

to form a new edge between two terrorists. Adding new edges, an Expanded Dual Graph adjacency was created. Afterwards, a comparison between original Dual Graph and Expanded Dual graph is performed.

A. Hidden Network Analysis

New formed edges were merged with the original TerroristRel dataset which has a shape of 851×1228. Even though we predicted that there should be an edge between those two terrorists, we still needed to figure out what kind of relationship that is and which label should be assigned. By using transductive learning method, we estimated the labels of these newly created relationships as family, congregate, contact or colleague. Here, there is no multi-labelling is allowed for new edges although we kept the initial multi-labelled relationships. In total, new 105 edges were added, so our new feature graph is a 956×1228 table. Comparison between the augmented network and original TerroristRel dataset is performed in terms of number of labels and presented in Table III.

Label	Orig. Edges #	Exp. Edges #	Increase
Colleague	461	513	% 11
Family	136	138	% 1
Congregate	106	144	% 35
Contact	148	157	% 6

TABLE III Percentage Change in Labels

Moreover, we carried out clustering for the giant component of the complete terrorist relations graph that we created by using K-Means. As input of the K-Means algorithm, $k = 4$ is given since we are expecting to see four different clusters that belong to particular labels. The comparison is demonstrated in Table IV. In Milestone 3, we have already clustered the original TerroristRel dataset. Nevertheless, we also performed the same algorithm on our original relationship network.

Labels	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Colleague	0.42	0.16	0.33	0.09
Family	0.35	0.06	0.00	0.59
Congregate	0.00	0.94	0.00	0.06
Contact	0.30	0.25	0.21	0.24

(a) Clustering Ratio of Original Dual Graph

Labels	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Colleague	0.00	0.33	0.37	0.30
Family	0.87	0.09	0.04	0.00
Congregate	0.00	1.00	0.00	0.00
Contact	0.00	0.57	0.28	0.16

(b) Clustering Ratio of Expanded Dual Graph

TABLE IV Clustering Comparison of Original & Expanded Dual Graph

B. Top Threat Heat Diffusion

Last but not the least, we performed a heat diffusion on our Dual Graph and Expanded Dual Graph. Heat diffusion is implemented with localized low-pass filtering of a graph with a dirac-delta signal on a particular node. Impulse signal is applied to one specific node of both dual graphs separately and comparison of the results is done visually. In order to choose which nodes we should put our impulse signal, we calculated different parameters for each node and selected potentially the most dangerous terrorists. First, we need to present the parameters and their specific interpretation for our case:

1) Degree: Number of the neighbors of every node is called degree. The most connected terrorist is simply a potential threat for society and its impact on the terrorist organization cannot be neglected. We ascertained that there are four highly connected terrorists and after them, there are several terrorists who have similar degrees. We strict ourselves to work with ten terrorists with highest degree for the purpose of choosing potential dirac-delta source nodes. These terrorists and their corresponding degrees are determined in the code.

2) Betweenness Centrality: Betweenness centrality of a "node v" is the sum of the fraction of all-pairs shortest paths that pass through "node v". The formula for betweenness centrality is:

$$c_B = \sum_{s,t \in V} \frac{\sigma_{s,t|v}}{\sigma_{s,t}} \tag{1}$$

where V is set of nodes, $\sigma_{s,t}$ is the number of shortest (s,t)-paths, $\sigma_{s,t|v}$ is the number of those paths passing through some node v other than s,t. We have computed shortest-path betweenness for every node and chosen the most central ten nodes for the purpose of choosing potential dirac-delta source nodes. This parameter can provide us the most *"wanted"* terrorist since it can be interpreted as the core of the operation. The Fig.2 is obtained regarding the mentioned importance. As it can be observed from the figure, two highest target terrorists should be Osama bin Laden (OBL) and Amar Makhulif, because they are the center of network.

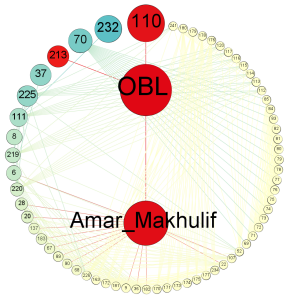


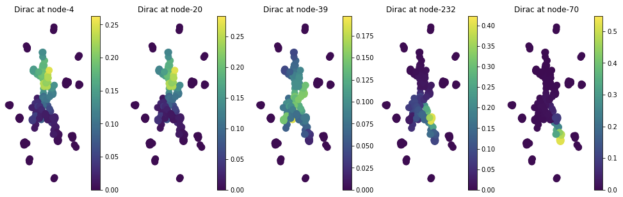
Fig. 2 Importance of Terrorist from Dual Graph

3) Articulation Points: An articulation point is any node whose removal (along with all its incident edges) increases the number of connected components of a graph. [3] We used NetworkX function that is dedicated for only this purpose. When we performed this function, we observed that there are many articulation points that create small additional connected components in addition to the giant connected component. We simply neglected those articulation points who create a second connected component with size smaller than 6 nodes. In that way, we filtered the unnecessary articulation points from our list and reached a list of important "key" terrorists. We suspect these terrorists may be crucial for a particular operation of a terrorist organization. Their role in the organization can be *"regional cell managers"* or even *contact person* with another terrorist organization. Removing a node from our dual graphs means removing a terrorist from the network. If that person is detained, we would isolate one component of the terrorist organization which is a strategically meaningful assault to carry out.

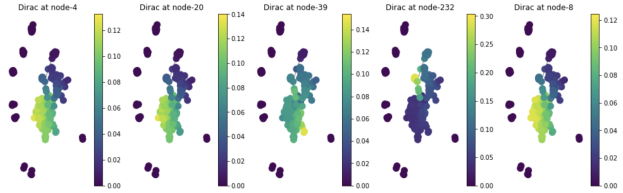
According to aforementioned parameters, we have chosen five terrorists for each graph that potentially constitute the biggest threat for the society. The chosen terrorists are presented in Table V.

Original Dual Graph		Expanded Dual Graph	
ID	**Name**	**ID**	**Name**
39	OBL	39	OBL
20	Mustafa_Kamel	20	Mustafa_Kamel
4	Amar_Makhulif	4	Amar_Makhulif
232	Zarqawi	232	Zarqawi
70	Abu_Khaled	209	Fateh_Kamel

TABLE V Chosen Terrorists for Heat Diffusion



(a) Original Dual Graph



(b) Expanded Dual Graph

Fig. 3 Heat Diffusion from chosen terrorist nodes on Dual Graphs

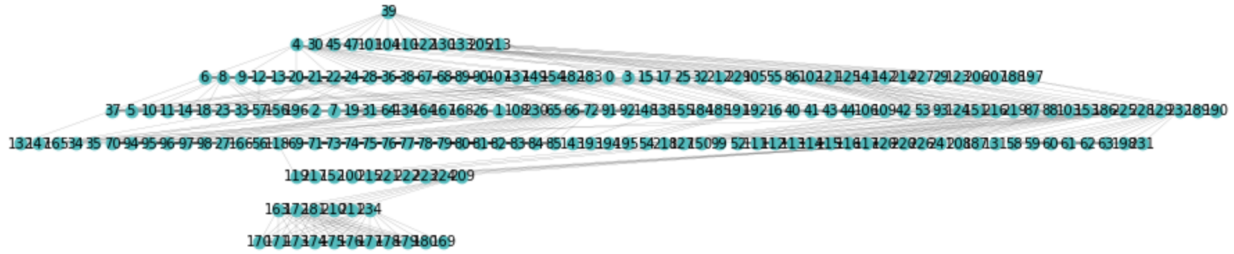


Fig. 4 Organizational Structure of the Giant Component of the Expanded Dual Graph. The head terrorist is verified as OBL with an ID of 39

IV. Conclusion

In summary, reconstructing the data has allowed us to look at the terrorist relations network from different perspectives. By creating the Dual Graph from the original relationship graph, we were able to predict the existence of relations that are not publicly known based on the similarity of node features. Reverting this expanded network back into a relationship graph format and learning the labels for the new relationships, we observed that this had a positive impact on the spectral clustering accuracy. Specifically, using spectral clustering, we managed to exclusively group the majority of family terrorist relations into a cluster. On the other hand, the Expanded Dual Graph allowed us to highlight the difference in threat between the observed and the *'true'* terrorist network. It also shed light on new terrorists that could be targeted compared to those terrorists that are central to the original network. Subsequently, we were able to validate the importance of a single terrorist (Osame Bin Laden) in both networks by analysing the spread of a signal emanated from his respective node in the network compared to other central nodes. Finally, reviewing his connections in the expanded terrorist network revealed an organisational structure (Figure 4) in his respective connected component that is comparable to a terrorist organization.

References

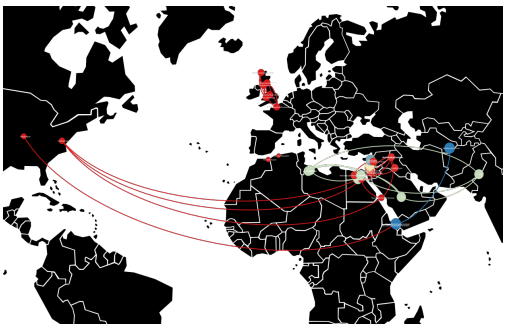
[1] Zhao B., Sen P., Getoor L.. *"Entity and Relationship Labeling in Affiliation Networks,"* Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[2] Doi E., *"Low-rank Decomposition and Logistic Regression Methods For Link Prediction in Terrorist Networks,"* CSE 293 MS Project Report, FALL 2010, University of California, San Diego.

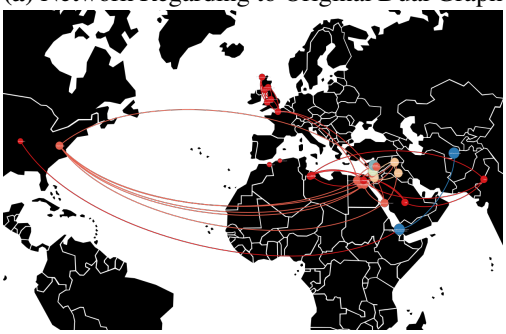
[3] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, *"Exploring network structure, dynamics, and function using NetworkX,"* Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.

APPENDIX

Additionally, we collected nationality of terrorists to visualize network in a geo-map. As these names doesn't appear in a single source, we checked manually for available information. We were able to find info for 35 terrorist out of 244. High proportion of these 35 terrorists are also the most important nodes regarding to explained betweennes centrality. The collected data is represented in the on the github repository. This 35 nodes network regarding to Original Dual Graph is represented in the Fig.5a and Fig.5b.



(a) Network Regarding to Original Dual Graph



(b) Network Regarding to Expanded Dual Graph

Fig. 5 Network of Known Terrorists. The size and the color of edges indicate modular communities formed by using Gephi.

While Fig.5a is containing 33 edges between 32 nodes, Fig.5b is containing 42 edges between 32 nodes. Also from the figures it can clearly seen that network became more complex. The most of the terrorists found by this search were affiliated to Al-Qaeda terrorist organization. As most of the Al-Qaeda members have Middle East origin, nodes are generally accumulated around Middle East countries.