Network Tour of Data Science Project : Finding the Authors of a Terrorist Attack

Yusi Zou, Loïc Nguyen, Maxime Lemarignier, Pedro Da Cunha
Team 25

***Abstract*—This report was written as part of the Network Tour of Data Science course taken at EPFL in 2018.**

The dataset we will use in this report is the Terrorist Attacks [1]. Our goal is explore and exploit the data to come up with a model to find the organization standing behind a terrorist attack for which no one claimed responsibility.

I. Introduction

In the last few years terrorism has made a notable return at the center of public attention. The recent events and the constant growth of tensions show us that this return is not going to end soon. This is why the comprehension of terrorist attacks and generally of terrorism is important. In this project we aim to discover more about terrorist attacks through data, to understand and to exploit them and finally to use it for good.

Our final goal is hence to build a model in order to predict which organization might stand behind a terrorist attack. Being able to quickly determine who might be responsible of an attack could potentially help security services in their investigations.

II. Project Structure

This project is composed of three main parts :

- **Data Acquisition** : We will first get to know our data. What does it contains? How is it represented? How can we shape it to make it usable in the context of our project? The main goal is to understand the data structure and to clean the data. We will also review various issues encountered with the raw data.
- **Data Exploration** : We will explore the data more in depth, observe what the corresponding network looks like and how the various parameters are distributed. We will also build a feature graph using the chosen techniques, and compute the similarities between the attacks to see whether it is possible to get a good result for our final goal.
- **Data Exploitation** : Now that we fully understand the data, it will be time to use this knowledge and exploit it. We will try to predict which organization stands behind a terrorist attack using different models, based on the knowledge of the features we have on this attack.

III. Data Acquisition

A. Raw Data Description

The dataset we use in this project is the TerrorAttacks dataset from the original Terrorist Attacks and Relations project. This dataset consists of 4 CSV files:

- The file *labels* contains the attacks labels, meaning the different types of terrorist attacks we have (Bombing, Kidnapping..).
- The file *nodes* present us the different nodes of the network. Each node represents an attack which is uniquely defined by an URL containing information of the organization behind it and the day it took place. Each one of these nodes is assigned a label and a vector of 106 binary "0-1" attributes, representing the presence or not of a feature.
- The file *loc.edges* contains the edges of the network. Two nodes are connected by an edge if they are geographically co-located.
- The file *loc_org.edges* is a subset of the previous file. It contains only the edges linking two attacks that are geographically co-located and made by the same organization.

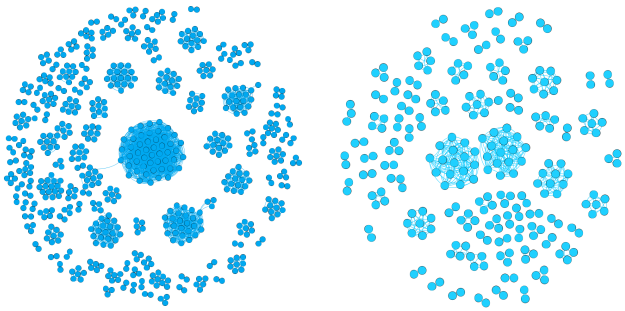
B. Building the graph

In order to build the graphs relevant to our project, the two provided edges files mentioned before are imported into Gephi to compute a visualization. The specialization used is *Force Atlas 2*. Its parameters were tweaked because all the connected components are cliques and the resulting graph was very dense. The gravity was set to 15 and the option to avoid nodes overlapping was used. The resulting graphs are shown in Figure 1, for *loc.edges* on the left side and the *loc_org.edges* on the right side. A more in depth analysis of them will be done later.

C. Processing the raw data

The data in our dataset is not displayed in a very readable way. Before being able to use it we need to process it.

In the TerrorAttack dataset, each attack is represented as an http URL. We split these URLs in order to create new columns, one containing the date of the attack and one containing the organization behind it. Some attacks do not have any organization linked to them. We hence attribute



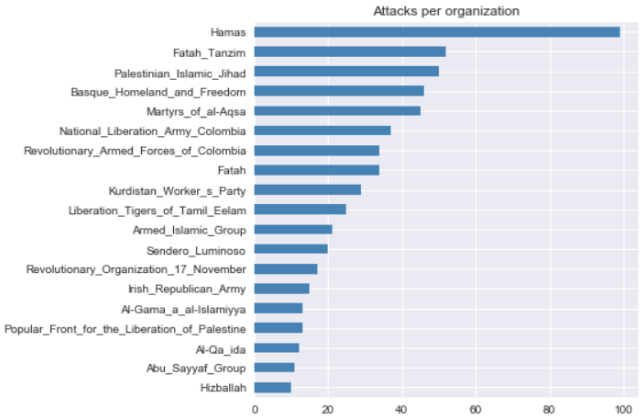


Figure 1. Graph on the left shows the the co-located attacks which are connected by edges. Graph on the right shows the co-located terrorist attacks performed by the same terrorist organization.

them the ”Unknown” organization. Information in these two columns will be of our main concern.

D. Notable issues with the raw data

We notice that almost half of the attacks do not have an organization linked to them. This shows that there is effectively an issue with determining who stands behind them and this is why we will try to predict them. It often happens that attacks are not claimed responsible by any organization. However an attack may also be the work of a single individual not affiliated with any group so we cannot always design an organization as responsible for it.

We also notice that in the form the data is given to us, if two attacks are conducted by an organization on the same day we have no way to differentiate them. Giving a unique id to each attack when creating the original data would have been a better model. Therefore we have no other choice but to assume that an organization conducts at most one attack in a day. Each attack is then uniquely defined by the organization and the day it was conducted.

The various attack nodes represented as URL links were pointing to pages that were maybe containing useful additional information. Unfortunately the website pointed by these links seems to be down so we cannot make any use of these links.

Finally we can also say that it would have been very useful to have the name of the various features for the attacks instead of integers. It would have allowed us to get a deeper understanding of the various aspects of terrorists attacks and helped us in our analysis.

IV. Data Exploration

Now that the data has been cleaned, it is a good idea to analyze its different statistics to really understand what's into the data.

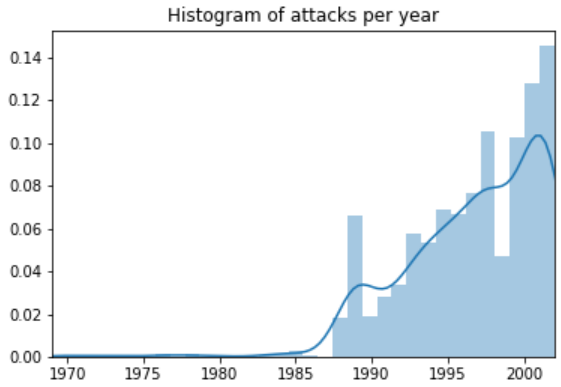
A. Attacks per organization

We start by having a look at the main organizations in the dataset. What are the most active organizations in terms of number of attacks?

We can see that ”Hamas” is by far the most active organization, with almost 100 of their attacks referenced in our data. We will use the organizations displayed in this chart in our model later in the project.

B. Attacks through time

How does the terrorism evolve through time? In order to answer to this question, we plot the frequency distribution of the attacks by year:



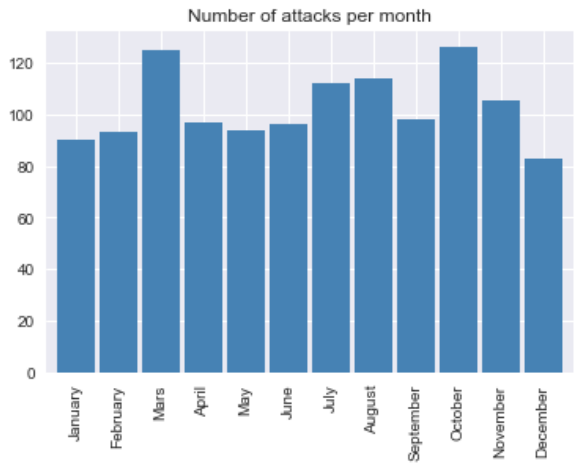
We see that the distribution of the attacks starts to rise somewhat linearly after the year 1958, with a peak around years 2000-2001. If we assume that the attacks referenced in this dataset were uniformly sampled from all the attacks that happened then we can observe that terrorism seems to be linearly expanding since the end of the 80s.

This is why in the scope of our project it might be interesting to be able to identify the organization behind an attack.

C. When do attacks happen in a year?

An interesting thing would be to see if attacks happen at random through time or not. If this is the case, we would expect to see a uniform distribution of the attacks. Otherwise we would ask ourselves whether we have periods of time

more prone to witness an attack. We now plot the number of attacks per month :



We observe that some months seem to witness more attacks than the others. For example, March and October have an increase of 47% in comparison to the month of December. However, we keep in mind that the data may be biased.

D. Building a feature graph

Because the *nodes* files contains a list of nodes with 107 features, it is a good idea to build a feature graph to analyze the data. We will be using the method seen during the tutorials [2]. When sparsifying the adjacency matrix, the treshold is set to 0.7.

The resulting feature graph is then generated in Gephi with the *Force Atlas 2* layout and is shown in Figure 2.

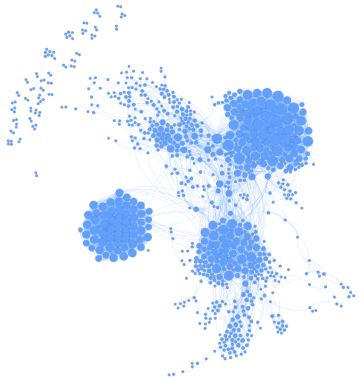


Figure 2. Feature graph

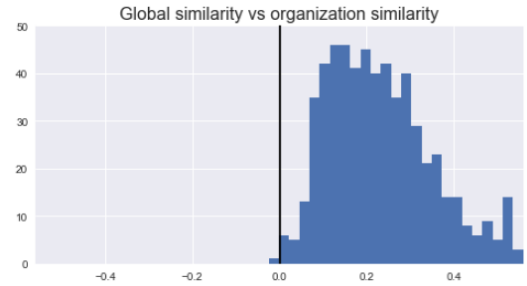
We see on the graph 3 main clusters that are strongly connected. Our hope is that these cliques represents attacks coming from the same organizations. For later uses, this graph will be sub-sampled because it is too complex as it is.

E. Attacks similarity

Since our goal is to predict the organization behind an attack we need to check if there exists relations between attacks of the same organization.

We first find the global similarity for each attach by computing the average cosine similarity in terms of features between this attack and all other attacks, then find the organization similarity by computing the average distance between this attack and all other attacks that are caused by the same organization.

We now plot a histogram of the values we find for organization similarity relatively to the global similarity.



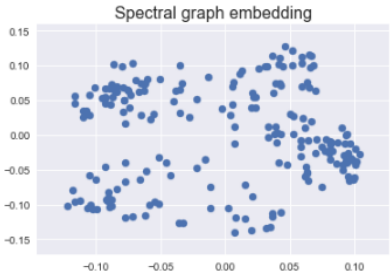
We observe that most of the differences are positive, which lead to the conclusion that there is indeed more similarity between attacks committed by the same organization than between two attacks taken at random. This implies that the presence or not of the features of an attack may give us information about its author. We can therefore try to apply a machine learning model on it and this is what we will now do.

V. Data Exploitation

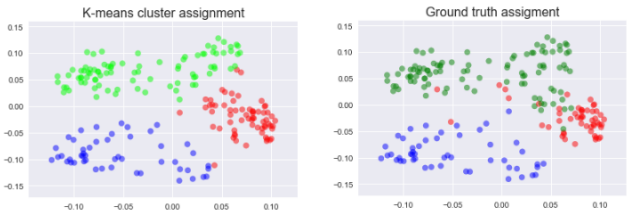
We now try in this part to exploit our data to see if we can predict the organizations responsible for each attack. The name of the organizations are used as labels. The first 2 sections are mainly used to see if attacks can be clustered based on its similarities. For both of these, we will only use the top 3 organizations that has the most attacks in our data (Hamas, Fatah Tanzim and Palestinian Islamic Jihad) because it will be easier to visualize. Each method will reduce our feature space into a 2 dimensional one where K-Means will be used to determine the clusters and make predictions.

A. Spectral graph embedding

The first method used to visualize our features data is spectral graph embedding. Since we decided to only take the top 3 organizations, the feature graph will be recomputed. With that feature graph, the normalized laplacian and its eigenvalues decomposition is computed and we embed the graph in 2 dimensions using the second and third eigenvector. As we can see, there are 3 distinct clusters. Our hope is that they may correspond to the 3 organizations we have.



From here on, the K-means algorithm is used to compute clusters assignments for each point.



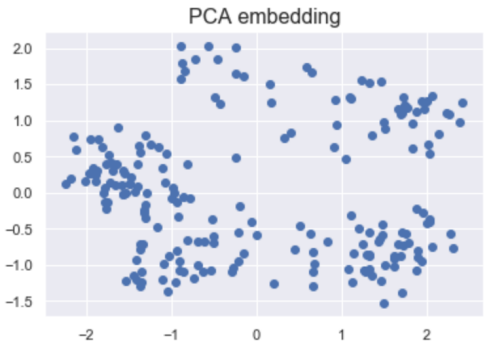
a) Cluster computed by K-means b) Real ground truth assignments

Figure 3. K-Means vs true labels

The graphs shows that the cluster assignments are very close to the true labels. The accuracy of our predictions is 88%. Therefore, it is a good indication that the data can probably be used to predict the organizations behind an attack.

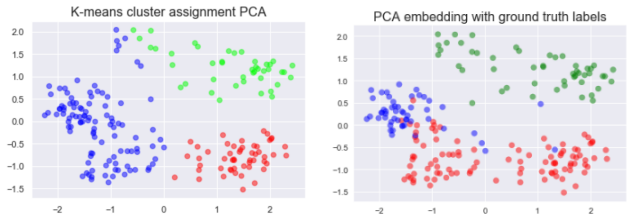
B. Principle Component Analysis

Another way to visualize is to use Principle Component Analysis (or PCA). This method is used to reduce high dimensional data into a lower dimension while keeping its most important properties. In our case, we will reduce our feature space to be able to plot it in 2 dimensions using the method from the *sklearn* library. [3]



We then use K-Means to compute clusters and find the accuracy of our predictions.

As we can see, the data indeed form some clusters. K-Means is now ran over the graph to predict the organizations. An accuracy of 71% is obtained. It is less than with spectral



a) K-Means cluster assignment b) Real ground truth assignments

embedding, but this again confirms our hypothesis that the data may be used to find organizations responsible for attacks.

C. Defining the Model

Since many organizations do not have a number of attributed attacks considered sufficient, we will only consider organizations with at least 10 attributed attacks. Hence our label set for the model will consist of 20 different organizations.

The features on which we will train our model are the attributes vectors of the nodes that represents the presence or not of fixed features in the given attack. The labels we want to give them are the corresponding organization which committed the attack. This is what we call a **multiclass label classifier**.

D. The Model

There doesn't exist many multiclass label classifier models. The one we chose to use for our problem is called *OneVersusRest* and we use its implementation from the *sklearn* python library [3].

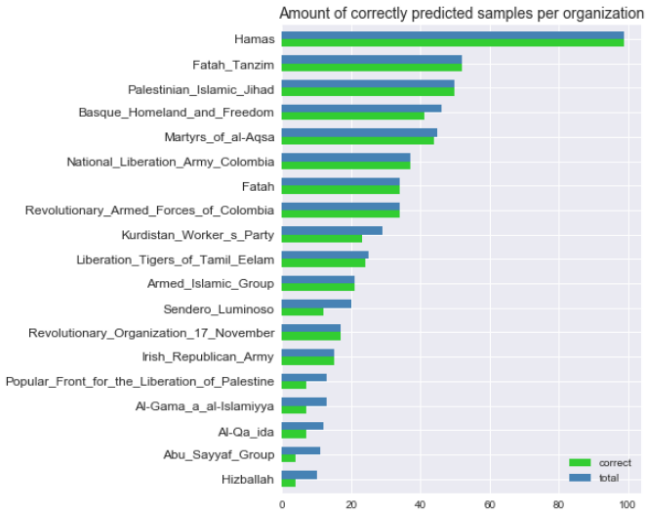
The OneVersusRest model is pretty simple. In our case it will train one classifier per organization, which will produce a confidence score that this attack is committed by this organization. For each prediction we have to make, we will apply all the classifiers on the attack and get their confidence scores. The model will then select the classifier with the highest confidence score and output the corresponding organization as the probable author of the attack.

E. Results

Here we display the results obtained with a 4-fold cross validation on our OneVersusRest model. The total number of attacks per organization is displayed in blue, followed by the attacks that were correctly predicted in green.

As we can see the model is doing pretty well. Our predictions are correct more than **91.2 %** of the time. We even manage to get perfect predictions for some organizations like "Hamas" or "Fatah". This hints that their attacks may have very distinguishable features that makes them easier to recognize from the rest.

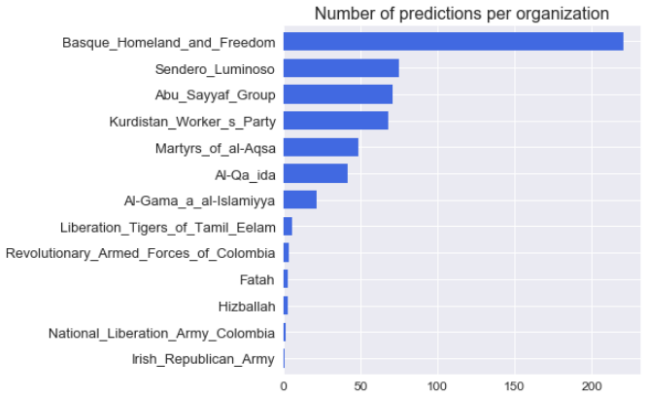
For some other groups however, specifically those which do not commit as much attacks we get worse predictions rates. Attacks committed by "Hizballah" for example are



misclassified more than half of the time. The lack of data could explain this, but smaller groups also tends to perform more various types of attacks in general, so this makes the task of correctly attributing their attacks to them harder.

F. Predicting for Non-Attributed attacks

Now that we have a working model for attributing the attacks we use it to try and predict the authors of all the non-attributed attacks in the dataset.



We observe that the vast majority of terrorists attacks are attributed to the ”Basque Homeland and Freedom” organization, followed by the ”Sendero Luminoso” and ”Abu Sayyaf Group” organizations.

The distribution of the prediction does not follows the training set, especially for the ”Basque Homeland and Freedom”, the numbers of the unknown attacks assigned to it is irregularly high. This may due to:

- The decision regions are precise for other organizations and for the rest of the data points which are unclear, they are all distributed to this organization.
- Some of the attacks are caused by individuals, and the features of these attacks are considered similar to the attacks committed by ”Basque Homeland and Freedom”.
- Error in the training step: the training error was not perfect for this organization.

G. Possible Improvements

Even if we achieved to get pretty good results with our model there is always room for some improvements.

First, our model would benefit from acquiring more fresh data. New organizations were created since year 2006 so we need to be aware of their existence to be able to attribute them attacks.

Another thing is that a great number of groups only have a few attacks referenced. If the group indeed does very few attacks, the chance that it did a given unknown attack is very little, but in case the group usually conducts more attacks but deny their responsibility of them, we will be less able to correctly identify the attack as theirs.

Also, if we were given the features names we could get a deeper understanding of what our model is doing. The importance of each feature being relative to the weight associated to it, we could interpret how each of these is important to the organization predictions. This could give us ideas for the generation of new features from the existing ones or could also help us understand which features do not really contain information and can be removed to reduce the noise.

Furthermore, if the sample size could be bigger, namely if we have more data points that represents attacks with a given organization, our model may predict better, as the amount of data is crucial for machine learning.

VI. Conclusion

We saw in this project that we could quite efficiently make a model to predict the organization behind a terrorist attack using our data. Some methods we tried perform better than others. The result demonstrates that the use of machine learning models in such domains could reveal itself useful for the investigations. Improvement are nevertheless still possible.

Terrorism is one of the biggest concerns nowadays. If we were able to discover the organization which stand behind unknown terrorist attacks, it may help to analyze the comportment and behaviour of the organizations, and combining information about international events and political analysis, it may also hopefully contribute to the world peace.

References

[1] “Terrorist attack dataset.” [Online]. Available: https://linqs-data.soe.ucsc.edu/public/lbc/TerrorAttack.tgz

[2] “Ntds 2018 tutorial. building a graph from features.” [Online]. Available: https://nbviewer.jupyter.org/github/mdeff/ntds_2018/blob/outputs/tutorials/02b_graph_from_features.ipynb

[3] “scikit-learn. machine learning in python.” [Online]. Available: https://scikit-learn.org/stable/index.html