A Network Tour of Data Science

A DATA STUDY OF TERRORISM AND ITS TENDENCIES

Zahra Farsijani, Joëlle Hanna
Dorsan Lepour, Amin Mekacher

https://github.com/AminMekacher/NTDS_Team19

Contents

1 Introduction

Terrorism is a hot topic in today's world. These recent years, this threat has become particularly acute, and the *Western society* has been more and more directly confronted with terrorism. Even though there is no universal agreement on the definition of terrorism, everyone knows the meaning of this term and everything it includes.[1]

The objective of this study is to analyze terrorism related data and extract specific trends and tendencies characterizing not only the attacks, but also the organizations and their modes of operation. Network science, spectral graph theory and graph signal processing are powerful tools which enable to build models and extract useful information.

This report presents the work accomplished as part of the course *A Network Tour of Data Science*, which gave the theoretical support behind these tools. The purpose of this study is **not** to make a **historical review** of events related to terrorism, neither a **geopolitical analysis** of areas it affects. It strives to use wisely the data science to answer **specific questions** and reach **conclusions**. Three distinct data sets were explored during the researches, and are presented in detail in the sections that follow.

2 Terrorist Relationships

2.1 Brief Description

This data set was created by the Mind Lab at University of Maryland. It contains information about terrorists individuals and their relationships. It is composed of 851 nodes, each one representing a **pair** of two terrorists. Each node has one or multiple label(s) which indicates a **relationship type** (colleague, congregate, family and contact). The edges link between them the nodes that contain a common individual among the pair. There are 8592 edges in total. The obtained graph structure can be seen as the **line graph** $\mathcal{L}(\mathcal{G})$ of a graph $\mathcal{G}$ where each node is a single individual (and not a pair), and the edges are their relationships. The difficulty with such graph $\mathcal{G}$ is that the labels are assigned to the edges and not to the nodes. The representation $\mathcal{L}(\mathcal{G})$ where the nodes are the adjacencies between edges of $\mathcal{G}$ is less intuitive but allows multi-label classification for relationships. Besides the label, each node has a series of 1124 attributes giving specific indications about the relation.

2.2 Limitations

The four previous milestones were all accomplished based on this data set and led to the following observations :

- The network is very sparse as the number of edges is very small compared to the maximum number of links, and the node average degree is 19.
- The graph is composed out of 13 connected components, the biggest one containing 665 nodes and having an average clustering coefficient of 0.0267.
- This giant component has some hubs, revealed by the presence of distinct clusters in the two dimensions embedding.

These emerging clusters disclose a particular structure of the network and are patently interesting to study. Diffusion behavior of the four different labels on the graph was indeed studied, but quickly showed limitations. Indeed, due to a lack of documentation[2] on the features and what they refer to, it had become of no interest to explore further this data set. One might think that it is possible to deduce the signification of some highly correlated features, but the huge number of them (more than the number of nodes themselves) and their binary values (indicating the presence of absence of a feature) led to the evidence that no reliable data processing could be applied on such scattered data set.

Some results and figures about diffusion and clustering were already given in the previous milestones and will not be shown here for conciseness.

[1] The U.S. Code defines it as the "use of violence or threat of violence in the pursuit of political, religious, ideological or social objectives", with an emphasis that it is "perpetrated against noncombatant targets".

[2] The website *http://www.mindswap.org/* which hosted the documentation and contained all the individual profiles of the terrorists and the features of the relationships is now closed.

3 Terrorist Attacks

3.1 Context and Introduction

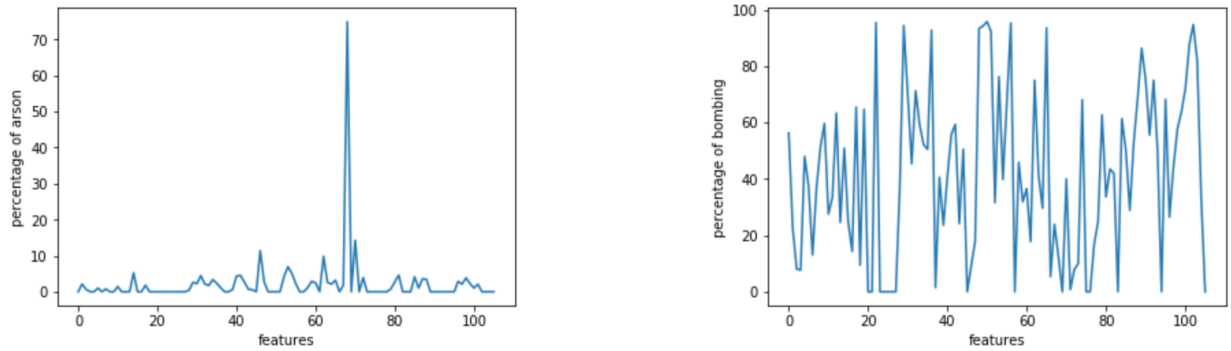
Given the blind alley encountered for the previous dataset, decision had been done to explore the other dataset provided by the same Mind Lab. This latter is composed out of 1293 nodes, each representing a terrorist attack. A unique label is assigned to each one and specify the operating means of this attack, among 7 categories (arson, bombing, weapon attack, kidnapping, NBCR[3], other). Two nodes are connected by an edge if the two attacks took place in the same region. The graph is composed of 787 connected components, the biggest one containing 51 nodes. Since edges link the attacks perpetrated in a common location, this essentially means that every connected component represents a unique region. A focus on the nodes and their single label can already divulge if certain region are dominated by a specific kind of attack.

In a similar way as before, a problem arises with the features associated to each node : it anew consists of a vector with only binary values. But this time, its size is 106 which is small compared to 1293 (the number of nodes). Compared to the previous dataset, where we had 851 nodes for 1124 attributes, it might now be possible to extract more information and try to interpret some of the features.

The achieved results are presented and explained in the next section.

3.2 Work Done

Firstly, we tried to understand data features (columns 1-106). Among various ideas we tried, we only resort to explaining one of them to be concise. We tried to understand if any of these features or their combinations could refer to a specific type of attack. We plotted the percentage of a type of attack (e.g. arson) versus the features. Two of the results are shown in figure 1. According to this figure, we conclude that it would not be possible to easily interpret the features and elucidate them.



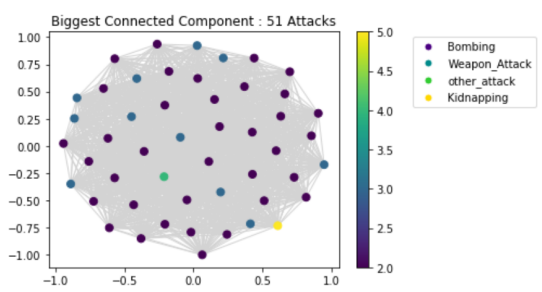
(a) Percentage of arson vs. features

(b) Percentage of bombing vs. features

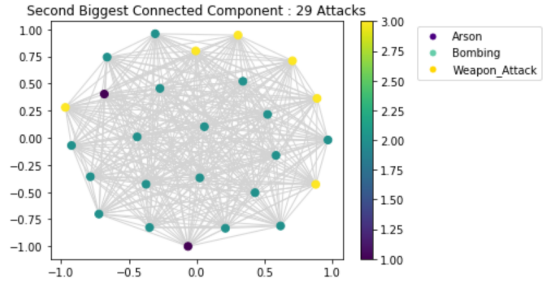
Figure 1: Percentage of two types attacks vs. feature vectors in largest connected component

Secondly, in order to understand the localization of attacks (i.e. if a location is dominated by a specific type of attack or not), we drew the labels on the four biggest connected components. The results obtained are summarized in figure 2. Since we have no documentation or further evidence to identify the regions (e.g. precise location details given by coordinates, names, etc.), we can not really make exact or relevant conclusions.

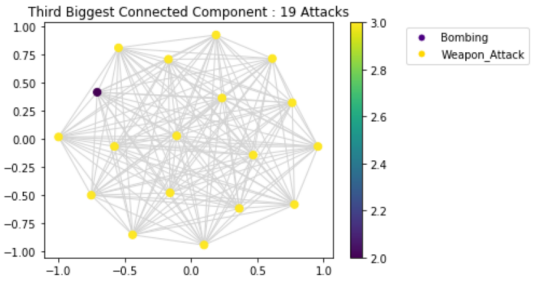
[3]Nuclear, Biological, Chemical and Radiological



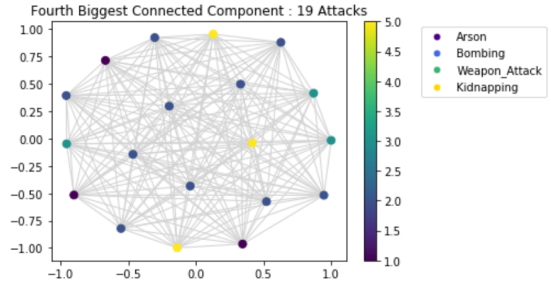
(a) Biggest connected component



(b) Second connected component



(c) Third connected component



(d) Fourth connected component

Figure 2: Labels on nodes of connected components

4 Global Terrorism Database

4.1 Context and Motivation

In order to have a more elaborate study of our problematic, i.e how terrorism movements have been able to grow in specific countries and their impact on some areas of the world, we decided to explore the Global Terrorism Database. Due to its important scale, we decided to down sample it, by only selecting the attacks happening during the last four years (2014 to 2017), which allowed us to study the recent trends in terrorism expansion. More precisely, we were interested in analyzing how the most vulnerable regions of the world, in our case Central Africa and the Middle East, were affected by such a threat gaining more dominion over their political, religious or economic independence.

4.2 Data Selection

By using the GTD Codebook[4] available online, we also selected the features we deemed useful to answer to our initial questions, which were mostly concerned with how the terrorist groups operate (the weapons they use, how their attacks are perpetrated) and how they select their targets. As such, the features we kept are the ones giving geographical information on the attacks (*county*, *province* and *city*); the weapons used and their sub-category if one is mentioned (*weaptype1* and *weapsubtype1*); the targets of the attack (*targtype1* and *targsubytype1*); the number of casualties both on the civilian side (*nkill*) and for the terrorists (*nkillter*); and finally, how the attacks were carried on (*attacktype1*).

4.3 Global Overview

In order to get an overview of the situation in both the Middle East and Central Africa, we used the Basemap library to display the node respectively to their longitude / latitude attributes. By using a color code, we discriminated the

[4]https://www.start.umd.edu/gtd/downloads/Codebook.pdf

attacks which are claimed by Boko Haram, ISIL, AQAP (Al-Qaida in the Arabian Peninsula) and the ones from other terrorist groups (or the unclaimed attacks). The result looks as follows (note: to have a manageable number of nodes, we only considered attacks that led to 10 or more civilian casualties):

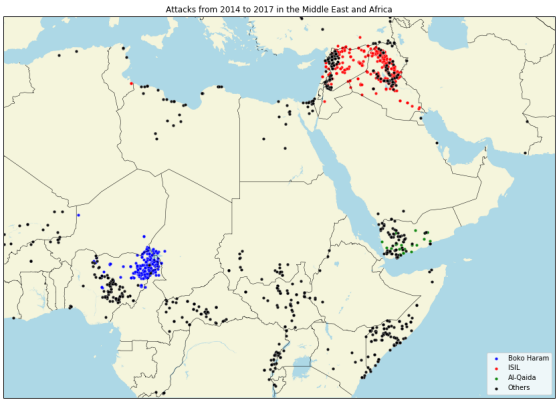


Figure 3: Attacks from 2014 to 2017 in the Middle East and Africa

We notice three hubs for the labeled nodes, respectively in Iraq for ISIL, in Nigeria for Boko Haram and a smaller one in Yemen for AQAP. We can then assume that the number of casualties will be much more consequent in these areas, as the aforementioned groups will have more leverage when it comes to move men and weapons around. To confirm this hypothesis, we plotted the number of casualties each country was subjected to during the four years we are studying. We can also compare this metric with the number of terrorists who died during these assaults in each country. By overlapping these two data, we get a good overview of how each terrorist group had to deal with international coalitions or other groups claiming their lands.

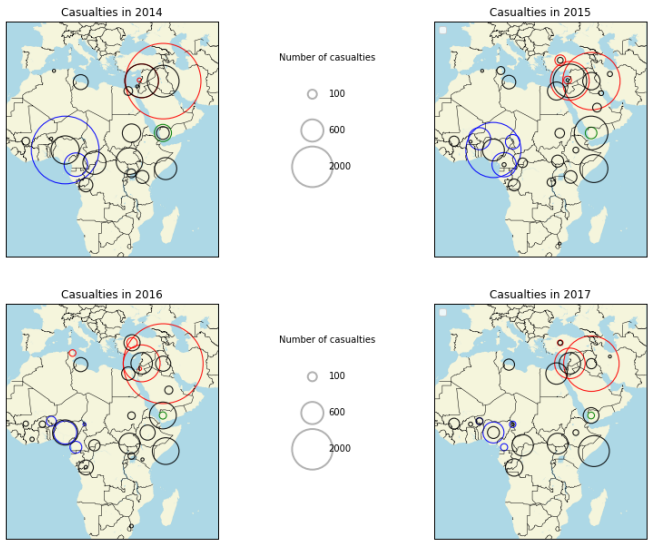


Figure 4: Number of casualties in each country for each year and each terrorist group. The color code matches the one used on the previous map

4.4 Features Analysis

As mentioned in the introduction, we selected some features in the original data set that we judged relevant for our research purpose. Therefore, we decided to study how each feature is distributed for each terrorist group. The goal of such an analysis was to see if there is a distinctive pattern for each group, for instance if their attacks are aimed towards a very specific target or if they have access to a large panel of weapons when carrying out their assaults.

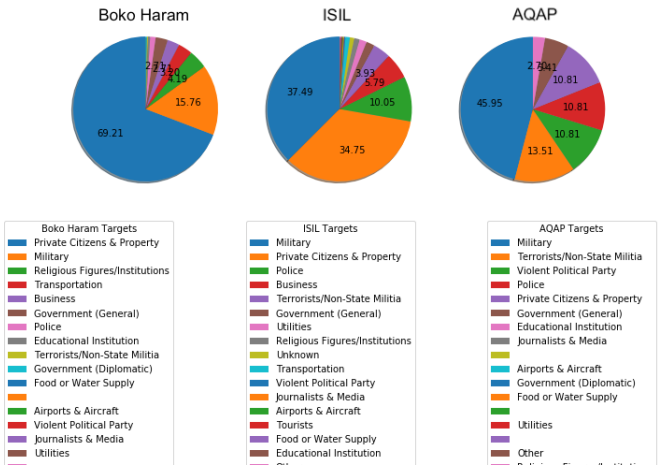


Figure 5: Targets distribution for each terrorist group

By using the probability of each feature, we can define a similarity weight between each classified node and the ones claimed by our groups of interest, in order to see if they are more strongly connected to one group in particular.

4.5 Features Classification

During the previous step, we stored the affinity of each unclassified node with every classified node, and now we are looking to see where this affinity is maximum for each node. If there is more than one node with the same affinity for the maximum value, we compare the group they belong to: we finally link the tested node with the group it is sharing the strongest similarity to.

In case one node has a very weak similarity with the three groups (less than 10%), we decided not to link it to any group: doing so avoids linking attacks which share a very small likeliness.

The initial map, with the new classified nodes displayed, looks as follows:

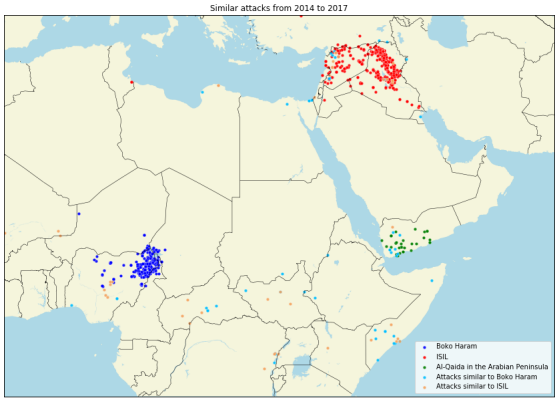


Figure 6: Attacks from 2014 to 2017 in the Middle East and Africa with similar attacks labeled