Predicting Terror Attacks?
A Data Story

Axel Nilsson, Nicolas Bollier, Elias Le Boudec, Enea Figini
Team 29

January 18, 2019

1 Introduction

Analysing terror organisations and predicting terror attacks is a subject of interest for national security organisations. From data on terrorist relationships and terror attacks, this project aims to assess whether terrorist relationships can be viewed as a social network, and to try to predict terror attacks locations from some known features.

To help reaching these goals, graph theory and data analysis tools provided by the course *A Network Tour Of Data Science* at EPFL will be used.

2 Exploring the Data
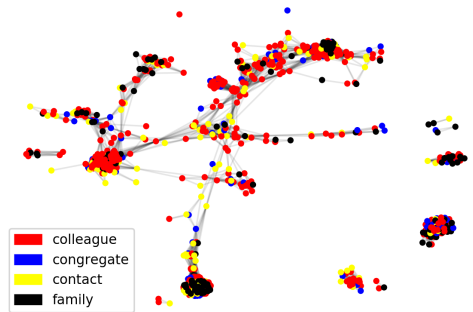
2.1 Relationships Dataset

Figure 1: *Terrorist relation dataset graph. The colouring of each node is related to the type of relation that corresponds to it*

This dataset is the line-graph of a network of relationship between terrorists. Each node represents a connection between two individuals and each edge an individual. The label of each note relates to the nature of the relation between the two individual. It is one of the following four: Family - congregate - colleague -contact. Two nodes are connected if they share one individual.

Our interest was to investigate the proprieties of the line graph as a relational network. As "original paper on dataset" mentions, an organisation need interpersonal connection to function and studying the structure of the social organisation could yield valuable insight. "Paper on line graph for social networks" found that on the basis of a study of an online social network, a social network could be well approximated by the line graph of a scale free network. If that propriety can be verified by our dataset, then we could gain information from the original graph from which the line graph originates.

Social sciences studies have shown that social/relationship networks have the particularities of homophily and transitivity. Logically if a & b are friends and c & b are also, the it is more likely that a & c are friends than not. This mathematically translates to:

$$a \sim b \text{ and } b \sim c \text{ then } a \sim c \tag{1}$$

As a first research question we will try to verify that our dataset derives from a scale-free network, implying that the graph that generated the relationship dataset have proprieties
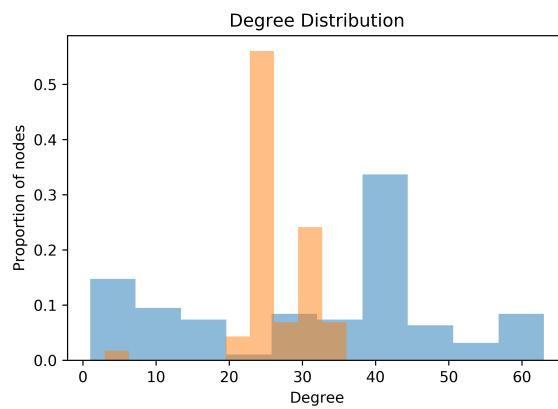
similar to social networks. By creating a scale-free network and making its line graph, we compared the degree distribution of the relationship dataset we were able to show that.
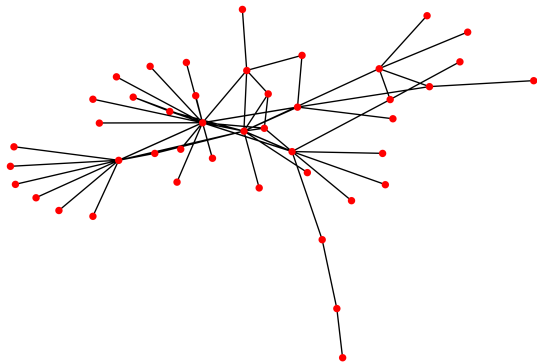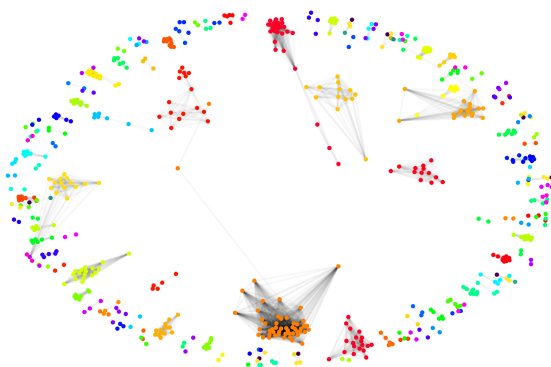

(a) *Scale free network*
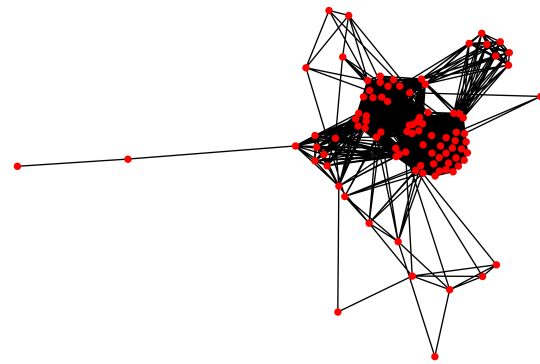
(b) *Line graph of scale free network*

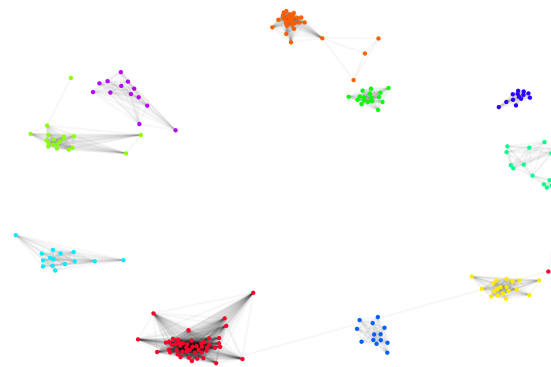(c) *Comparison of the degree distribution of the two line graphs. In blue, the created line graph and in orange the degrees of the relationships graph*

Figure 2: *Graphs from the terror attacks dataset*

2.2 Terror Attacks Dataset

(a) *Terror attacks location graph, colouring by component ID*

(b) *Ten biggest components from the terror attacks location graph*

Figure 3: *Graphs from the terror attacks dataset*

Similarly to relationships datasets, we have found that the terrorist attacks dataset have very highly connected components, implying transitivity. If attack a took place close to b, and attack b took place close to c, then it is probable that attack a took place close to c.

3 Comments on the Data

3.1 Terrorist Relations Dataset

3.2 Terror Attacks Dataset

Multiple issues have been found in this dataset:

Broadness   The dataset comprises attacks ranging from 1969 to 2005 and spanning the entire globe. Simple and relevant explanations for the graph formation or properties are not likely to be found, since the mechanisms behind two different attacks can be entirely different.

Structure   Half of the nodes are isolated, hence the topological information they carry in the graph is very limited. What is more, because of the transitivity relation described in Section 2.2, connected components are in most of the cases complete, hence isotropic.

Reliability   Errors have been found in the data. For example nodes `Djibouti_Youth_-Movement_19900927` and `Armed_Islamic_Group_19950711` have been connected, whereas the first attack took place in Djibouti [1] and the second one in Paris [2]. Hence algorithms using the data must tolerate some error in order to avoid overfitting.

Incompleteness   The dataset has been constructed from publicly available sources [3]. Because of the sensitivity of the data behind terrorist attacks and relationships, some of it is classified, making the dataset incomplete.

4 Predictions

The algorithm used to predict the terror attack location is the following:

1. From the dataset, select the 10 biggest connected components ("component" in what follows).

2. Sort the dataset by date of terror attack.

3. At this point, a component represents a location, and the nodes are the terror attacks in chronological order.

4. Select one node per component that is strongly connected to the others, the "lead" node.

5. Find the lead node $l^\star$ that is the most strongly linked to the new node (i.e. the next terror attack).

6. The predicted location of the next terror attack is the location of the component $l^\star$ belongs to.

The determination of the lead node uses the features vector supplied with each node, and a weighting function w. Let w be the application that returns a weight for each pair of nodes (n_1, n_2) in the graph $\mathcal{G}$, defined as

$$w : \mathcal{G}^2 \to \mathbb{R}^+ \tag{2}$$

$$(n_1, n_2) \mapsto f(|n_1 - n_2|) \tag{3}$$

where

$$|n_1 - n_2| = \|\text{features}(n_1) - \text{features}(n_2)\|_2 \tag{4}$$

features(n) is a binary features vector for each node n in $\mathcal{G}$ and $f : \mathbb{R}^+ \to \mathbb{R}^+$ is the node distance weighting. Examples for f are given in Table 1.

For each connected component, the lead node is determined as described below.

Algorithm 1: *Finding the lead node of a connected component with weighted edges*

Data: Connected component C
Result: Lead node n_l
Initialise $s(n)$ to zero. s is a dictionary mapping a score $s(n)$ for each node n
for *each edge e from C* **do**
    Let $e = (n_1, n_2)$, w be the weight of e
    $s(n_1) \leftarrow s(n_1) + w$
    $s(n_2) \leftarrow s(n_2) + w$
end
return $n_l = \arg\max_{n \in C} s(n)$

Finally, the prediction algorithm is presented below.

Algorithm 2: *Finding the predicted location of the next terror attack*

Data: Set of connected components $\{C_i^t\}$, $i = 1, \ldots, 10$, and the features vector of the next terror attack n_{t+1}, i.e. features(n_{t+1}), at each timestep t
Result: Location prediction p_t for time $t+1$ at time t, at each timestep t
for *each timestep t* **do**
    Compute the lead component $l(C_i^t)$ for each component C_i^t
    $p_t = \arg\max_{i=1,\ldots,10} w(n_{t+1}, l(C_i^t))$
end

4.1 Justification

The design of prediction algorithm is motivated by the following aspects:

– The labels are taken into account by weighting the edges. This allows to completely ignore label signals on the graph and simplify the analysis.

– The determination of one lead node per component allows to smoothen local variations inside a component, thus making the prediction algorithm more robust.

– The choice of one lead component per component is justified by the fact that connected components are almost complete.

4.2 Results

Table 1: *Prediction accuracy for different node distance weightings f*

Weighting		Best skewness ζ	Accuracy
Gaussian:	$f(d) = e^{-d^2/\zeta} - e^{-1/\zeta}$	0.01	50.5 %
Log-Exponential:	$f(d) = e^{-d} \log\left(\frac{1+\zeta}{d+\zeta}\right)$	0.1	50 %
Linear:	$f(d) = 1 - d$	N.A.	47 %
Square:	$f(d) = \begin{cases} 1 & d < \zeta \\ 0 & \text{otherwise} \end{cases}$	0.1	43 %

5 Conclusion

Analysing the terrorist relationships graphs showed that the network does not correspond to a scale-free network, which is typical for social networks.
The prediction algorithm yields reasonable results, given the flaws contained in the data and the compromises that had to be made (removal of isolated nodes, selection of a lead node).

References

[1] Amnesty Internation Publications, 1 Easton Street, London, *Amnesty Internation Report 1991*, 1991.

[2] L'Obs, "Attentats de 1995 : chronologie." [fr] Online. https://bit.ly/2ASwNQP, last checked 17 January 2019, October 2007.

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