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

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

Guideline

Data Acquisition: dataset; data cleaning

Data Exploration: network; distribution of parameters; feature graphs; possibility of doing machine learning to predict

Data Exploitation: finding of the best model; result of prediction

Possible improvement and conclusion

I - Data Acquisition

Raw Data

4 csv files :

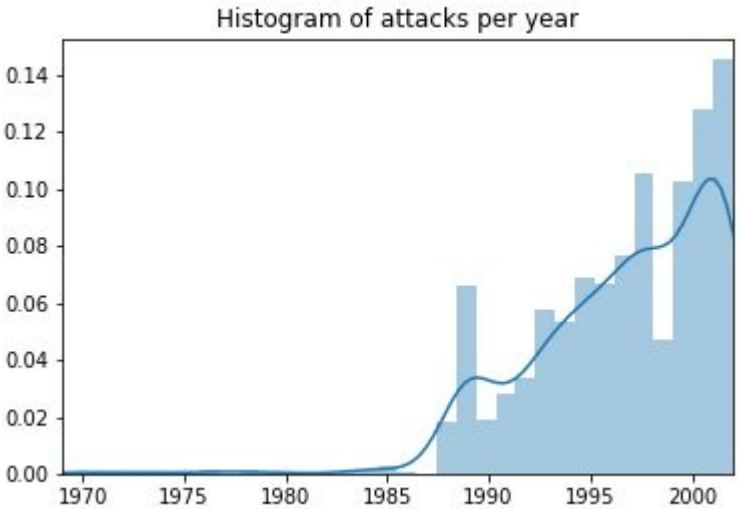
- Attack labels
- Attack **Nodes** with features
- **Edges** of co-located Attacks
- Edges of co-located Attacks by the same organization

Pre-Processing the Data

- URL -> organization + date
- Missing data, empty attributes
- No names for the features, only numbers

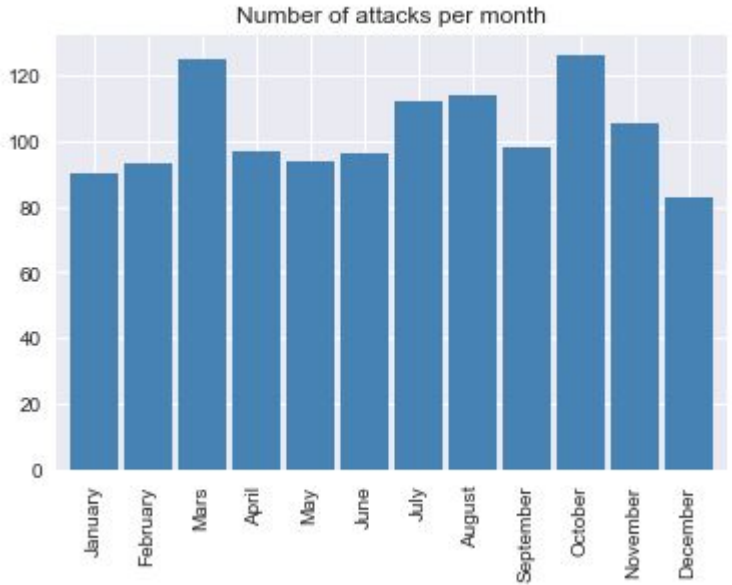
II - Data Exploration

Evolution of Attacks through time



Almost linear growth of number of attacks in the last few decades, assuming data is sampled uniformly.

When do Attacks happen ?

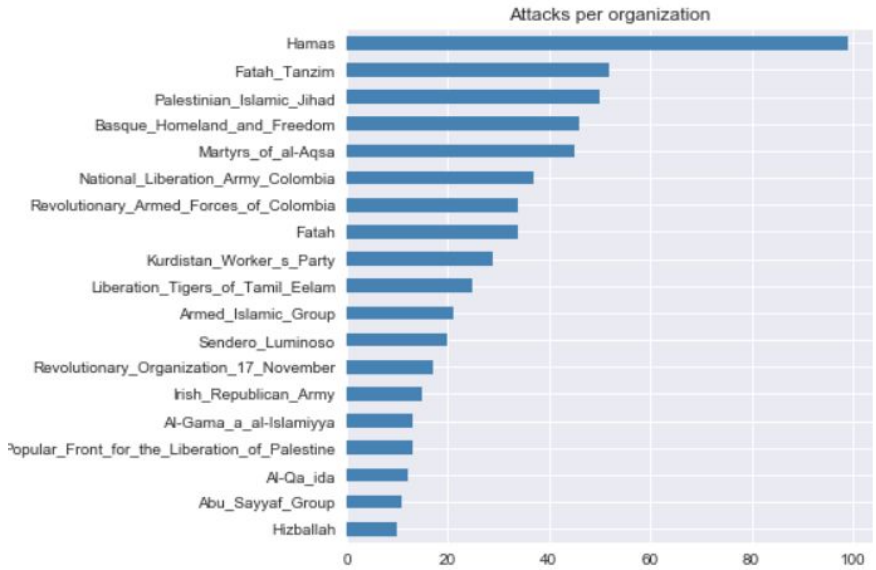


Some months seems to witness more attacks than others.

47% more attacks in October than in December !

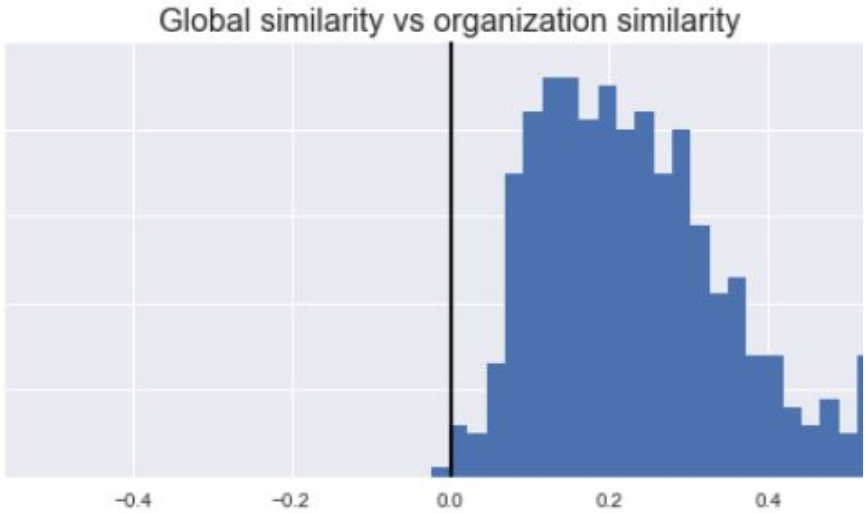
More data could show if this is a bias or not.

Attacks per organizations



- Hamas : ~ 100 attacks
- 80.3% of the known attacks are made by frequent organizations.

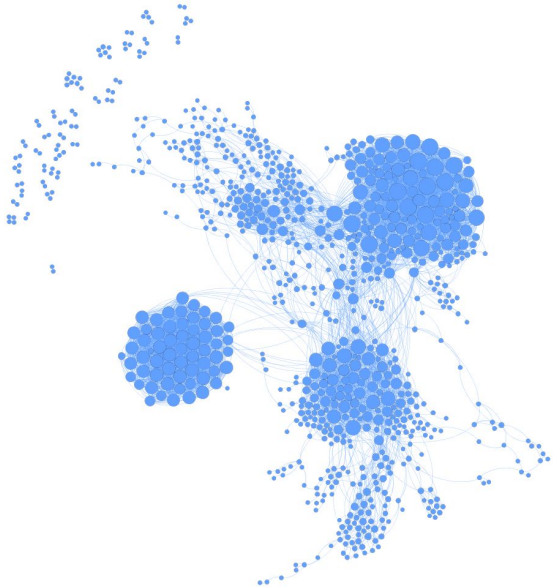
Attacks Similarity



Cosine similarity of the features.

Attacks committed by the same organization are on average more similar than between random attacks.

Building a Feature Graph



- Transform the types of attacks (given labels) into features using dummy variables encoding (adds 6 more features)
- Compute the feature adjacency matrix using Gaussian function and euclidean distances
- Sparsify the matrix (threshold 0.7)
- Visualize in Gephi

III - Data Exploitation

The Problem

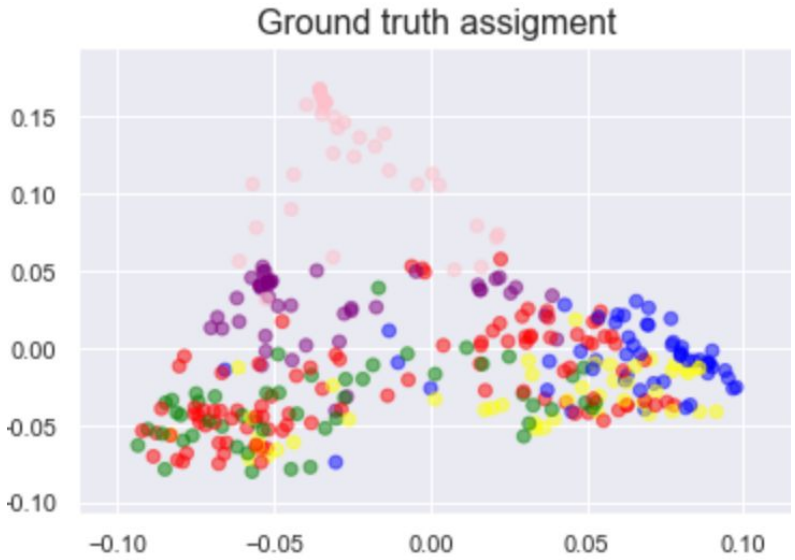
- More than half the attacks in our dataset is not associated with any organization
- Attacks often not claimed by anybody
- Important issue in the context of investigations

We want to build a model to accurately predict the organization behind an attack.

Spectral Graph Embedding

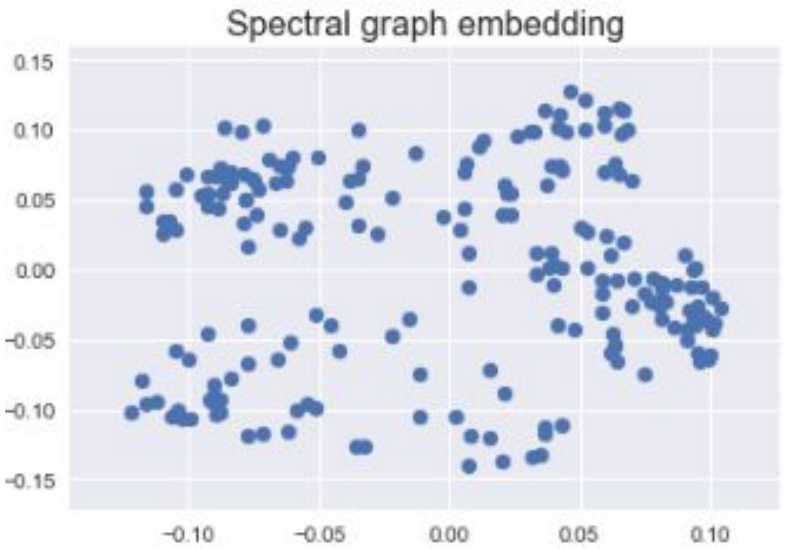
- Compute the laplacian and the eigenvalues decomposition
- Embed the graph in 2 dimensions
- Run K-Means to find out predictions

Spectral Graph Embedding



- Not so good with many labels…
- K-Means won’t give good predictions

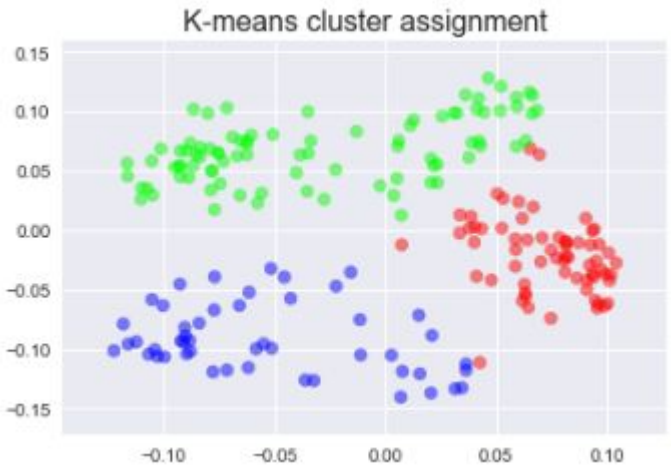
Spectral Graph Embedding



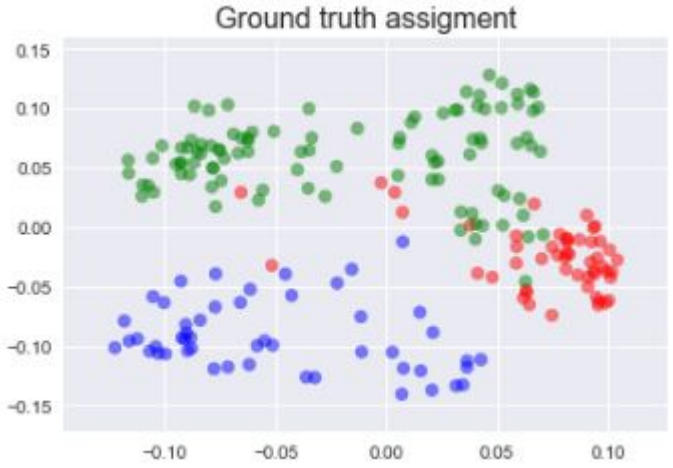
- Take only the top 3 organizations with most attacks (Hamas, Fatah Tanzim and Palestinian Islamic Jihad)
- Recompute the feature graph and embedding

Spectral Graph Embedding

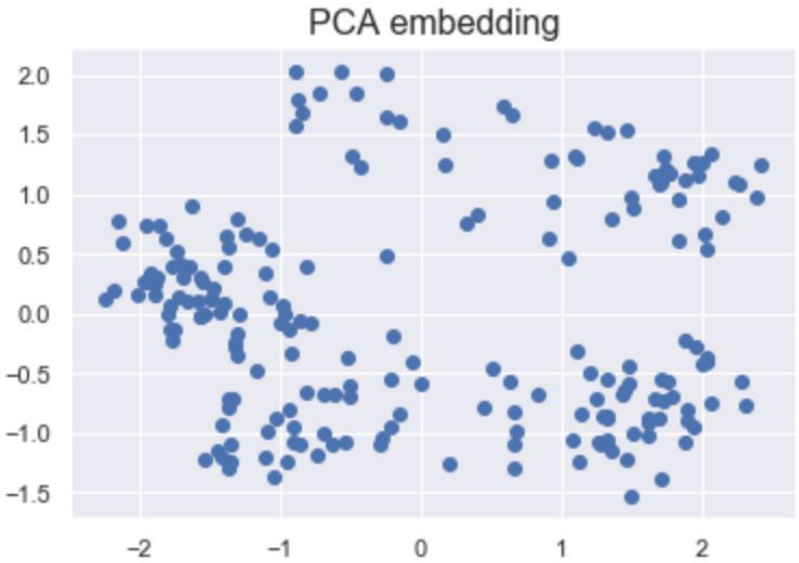
K-means vs true labels



88% accuracy



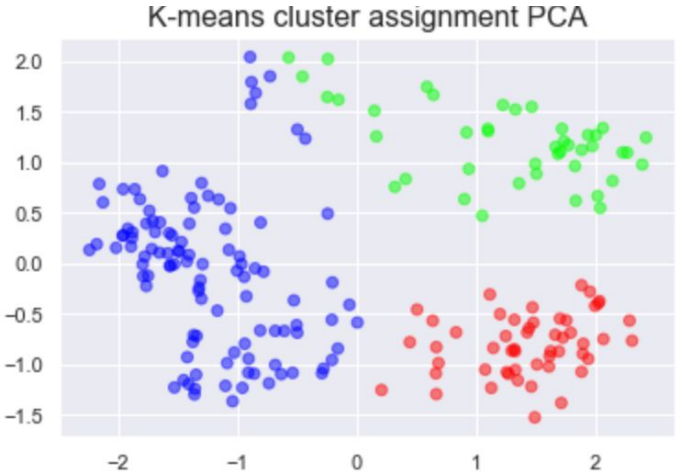
Principal Component Analysis



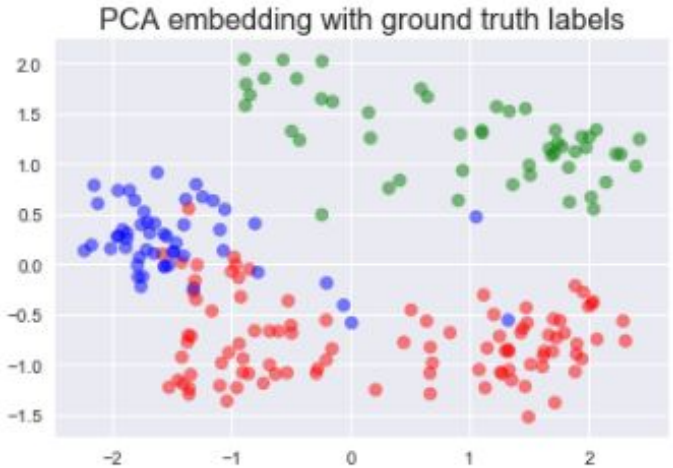
- Dimensionality reduction on the features based on Singular Value Decomposition
- From 113 to 2
- Visualize it on a 2D plot
- Run K-Means to make predictions

Principal Component Analysis

K-means vs true labels



71% accuracy



The Model

We saw that features contains information about the organization.

Use the features of an attack to determine the organization behind it.

Multiple label organizations, we need a multiclass label classifier.

We use the sklearn model named **OneVersusRest**.

OneVersusRest Model

Goal : Output most probable organization

- Train one classifier per organization that outputs a confidence score
- For an attack run all classifiers
- Output the organization with highest confidence score

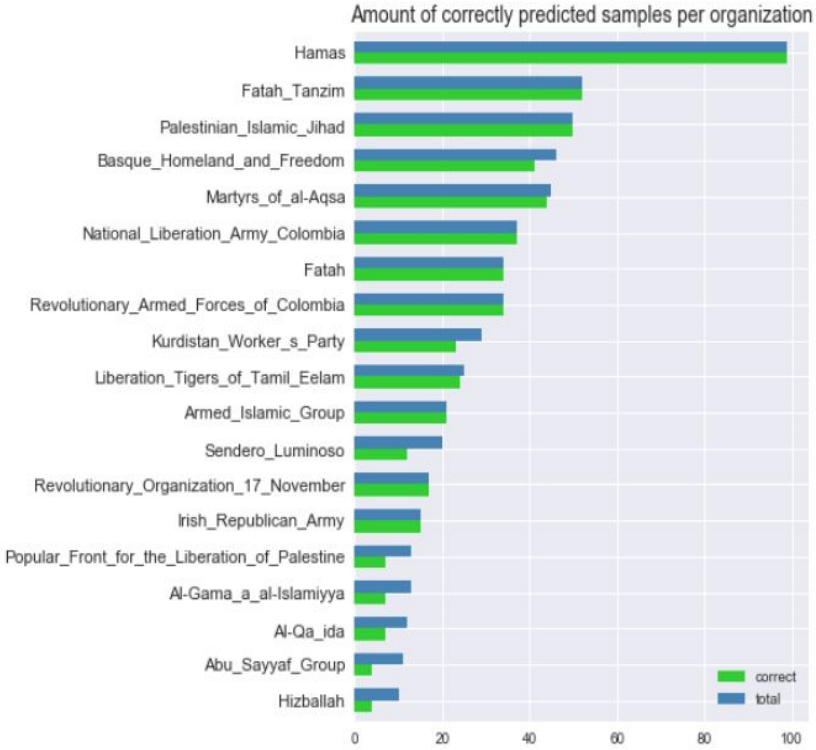
Evaluating the model

Cross-validation on known organizations.

Correct predictions 91.2% of the time !

Some organizations perfectly predicted.

Harder to predict for smaller organizations or with less attacks.

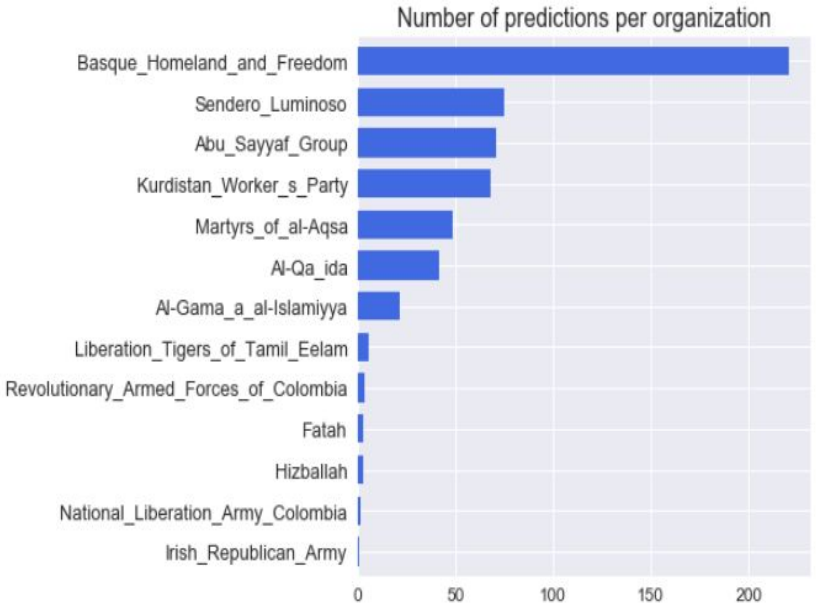


Predicting the attacks

Few of the main organizations are represented.

A bit suspicious. Actually good predictions? Or unknown attacks unrelated with all organizations?

May be the truth but no way to verify that.



Possible improvements and conclusion

Possible Improvements

We saw that a model for predicting organizations is possible.

Still many possible improvements :

- Name of the features : deeper interpretation, feature generation
- More recent data : new organizations and attacks since 2002
- More training data

Conclusion

OneVersusRest model surpassing K-Means on spectral graph embedding and PCA.

The model is seemingly good but we have no actual way to verify the predictions.

Nevertheless, this study shows the potential of machine learning for investigation of terrorist attacks.