UNSUPERVISED ANOMALY DETECTION WITH ISOLATION FOREST

WITH APPLICATION TO TRANSACTIONS MONITORING FOR AML

PyData London 2018

Speaker: Elena Sharova

TALK OUTLINE

- Introduction: What is Anomaly Detection and What Methods are Used?
- Applications of "Screening for Unusual Activity" in Banking and Why it is Difficult
- Supervised vs. Unsupervised Methods for AML
- Description of Isolation Forest (IF)
- IF Implementation in scikit-learn, Compared Performance to LOF on KDDCUP99
- IF Performance on Synthetic Dataset – Subspace Outlier Detection (AML Red Flags)
- Conclusion: Ask the Right Questions and Choose the Right Data Model

Disclaimer

Everything said in this presentation and the content of these slides is based solely on the author's opinion and research and bears no relation to my current employer.

I am not authorised to speak on my employer's behalf or disclose any work I perform for my employer.

INTRODUCTION

Anomaly=Outlier=Deviant or Unusual Data Point

When data generating process behaves unusually it results in outliers.

The subject of outlier detection is a well-researched area and there is sufficient amount of literature that covers it in statistical and data science.

Often, the real challenge in anomaly detection is to construct the right data model to separate outliers from noise and normal data.

METHODS

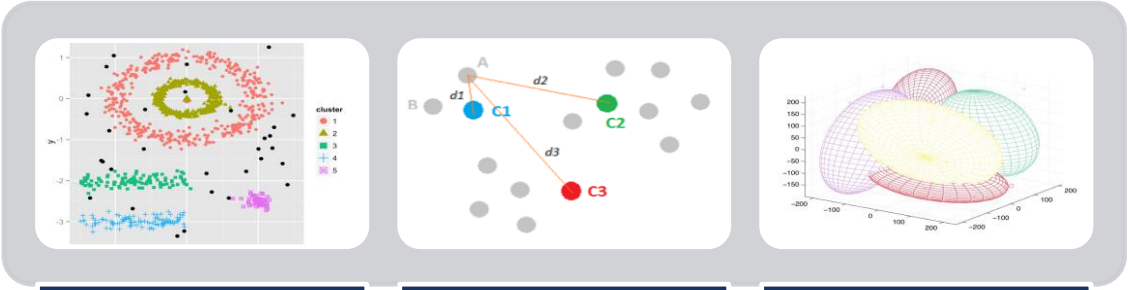
Density-based:
- DBSCAN
- LOF

Distance-based:
- K-NN
- K-MEANS
- Regression hyperplane distance

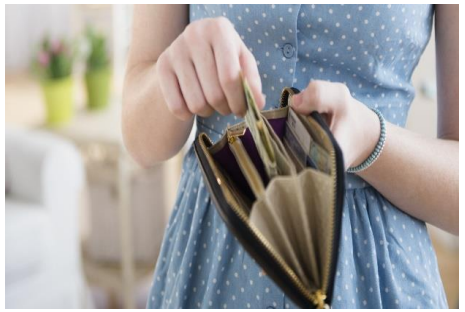
Spatial Proximity

Parametric:
- GMM
- Single Class SVMs
- Extreme value theory

Other: statistical tests (e.g. Z-score), variations on the above



APPLICATION TO BANKING

Retail Bank	Private Bank	Investment Bank
		
- Credit card fraud - ML through Retail Bank	- Market abuse - ML through Private Bank - Other fraud	- Market abuse - ML through Investment Bank - Other fraud

Done in Different Ways, Requires Different Approaches Because RED FLAGs are banking-type specific (especially the layering/structuring stage).

SUPERVISED VS. UNSUPERVISED

Automatic ML Detection is inherently a different problem from Automatic Detection of Credit Card Fraud and Market Abuse.

- In Automated Credit Card Fraud detection we know what a TP looks like. How? Customers tell us. Can use supervised approach because the true class is self-revealing.
- In Market Abuse detection we often know what a TP look like. How? Positive PnL and price moves can indicate this. Also self-revealing.
- In Automatic Money Laundering detection we don't really know whether a data record is a TP or not. A predictive model for Money Laundering is almost uniquely an UNSUPERVISED learning problem.

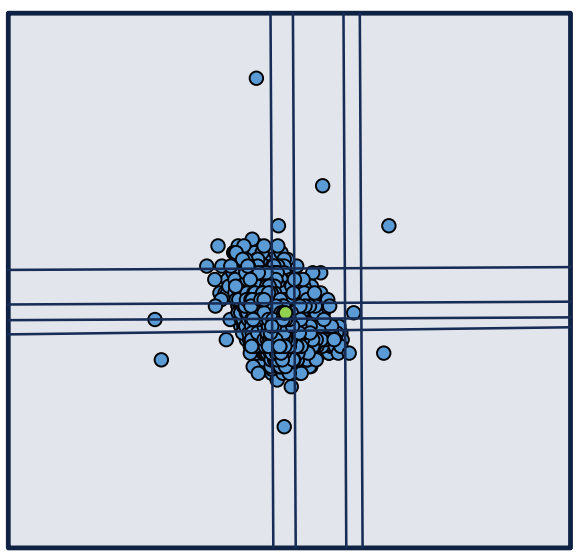
WHY IS UNSUPERVISED DIFFICULT

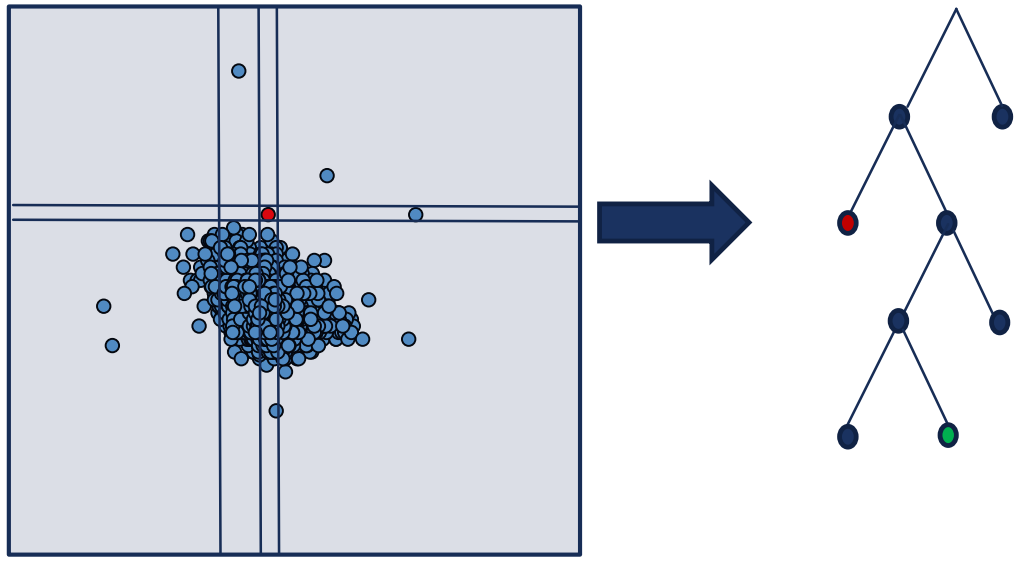
The main issues for an analytical approach to AML are [2][5]:

- SEVERE CLASS IMBALANCE – estimated less than 0.1% of wires a day in US involve money laundering.
- SEVER CLASS OVERLAP – money laundering is mixed with legal financial activity, especially in IB.
- CONCEPT DRIFT – how money is laundered evolves and changes all the time even by the same criminal organisation.
- UNCERTAINTY AROUND THE DATA MODEL (in addition to complexity and volume of data).

ISOLATION FOREST

Isolation Forest [3] is an ensemble regressor, and it uses the concept of isolation to "explain/separate-away" anomalies.
No profiling of normal instances, and no point-based distance calculation. Instead, IF builds an ensemble of random trees for a given data set, and anomalies are points with the shortest average path length.





ISOLATION FOREST CONT.

IF can work as supervised and unsupervised classifier.

IF calculates an *anomaly score* = $2^{-\frac{E(h(x))}{c(n)}}$ where h(x) is the number of edges in a tree for a point x, and c(n) is normalisation constant for a data set of size n.

In supervised setting, a threshold on an anomaly score separates binary class. The same anomaly score can be used without a threshold, thus allowing for a soft-classifier in unsupervised setting.

As $E(h(x)) \rightarrow 0$, anomaly score $\rightarrow 1$

IMPLEMENTATION IN SCIKIT-LEARN

Isolation Forest (IF) was introduced in 2008, and became available in scikit-learn v0.18 in 2016.

IF extends BaseBagging regressor (Bootstrap Aggregated Regressor), and it is possible to control bootstrap parameter (True=will replacement, False=without).

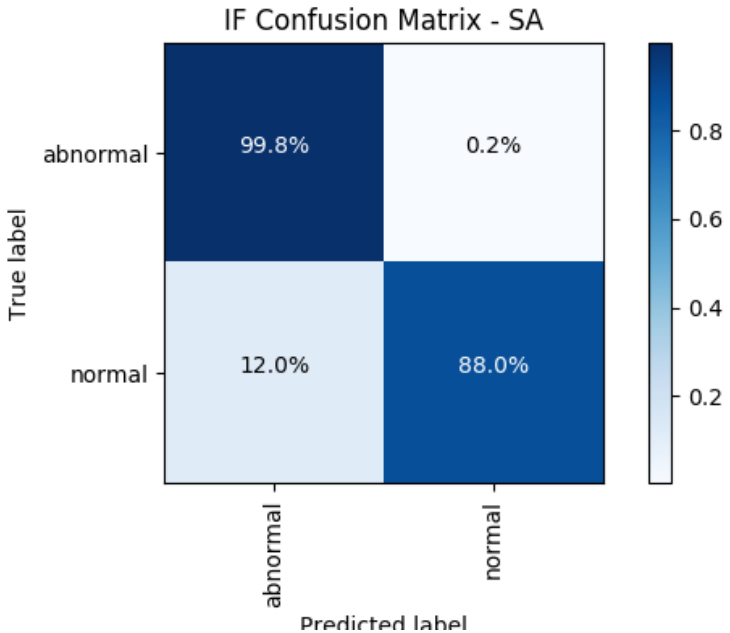
Its base estimator is ExtraTreeRegressor – an extremely randomized tree regressor. It splits on the best split among randomly chosen attributes with randomly chosen split points. This helps to overcome overfitting and locally-greedy trees.

It returns (0.5 – anomaly score), thus smaller scores are attributed to anomalies.

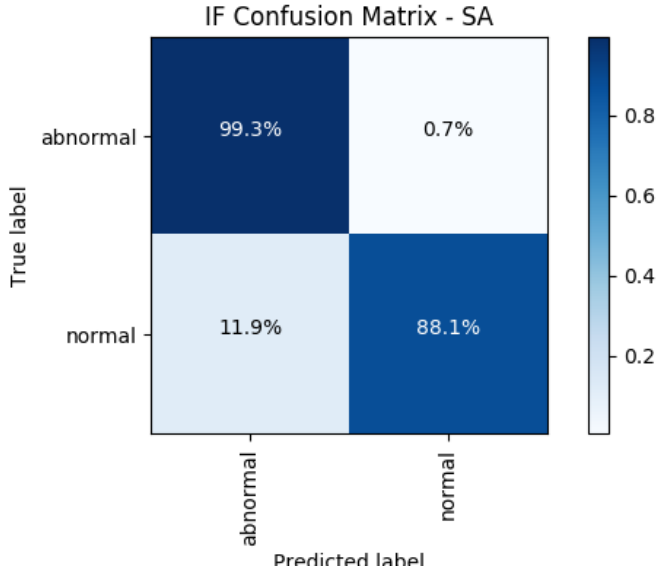
PERFORMANCE ON KDDCUP99

KDDCUP99 is a public datasets of logs from an off-line intrusion detection system. SA has 41 attributes, SF has 4. Link to Jupyter notebook: *https://github.com/elena-sharova/IsolationForest/blob/master/IsolationForest_v0.1.ipynb*

Training – AUC 93.7%

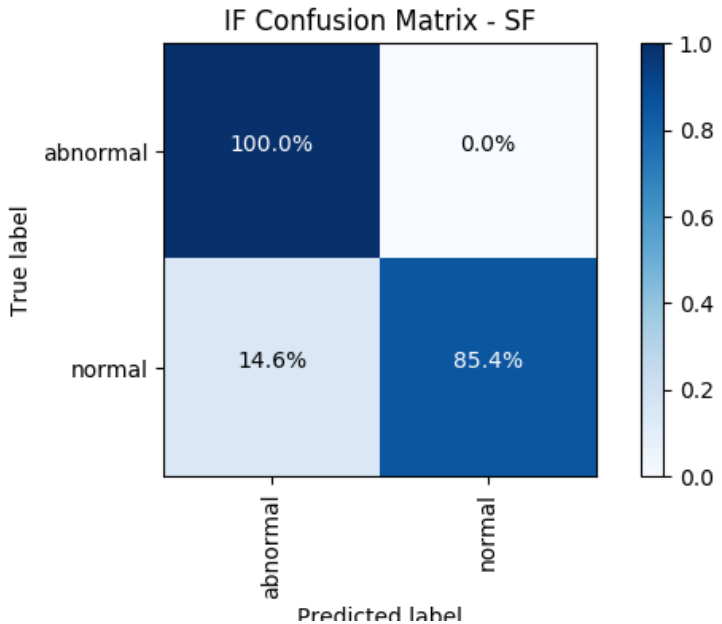


Testing – AUC 93.7%

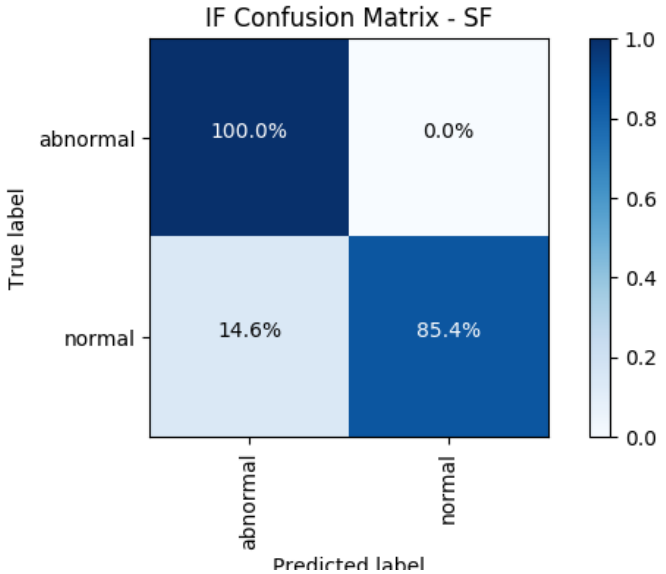


PERFORMANCE ON KDDCUP99 CONT.

Training – AUC 92.7%



Testing – AUC 92.7%



PARAMETERISATION

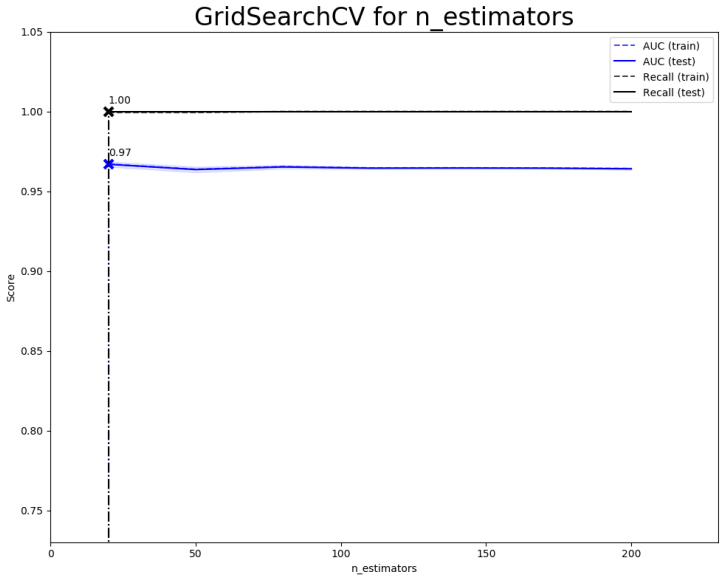
IF has two stages: training and testing/score assignment.
Training stage involves building iForest (from iTrees) and testing stage involves passing each data point through each tree to calculate average number of edges required to reach an external node.

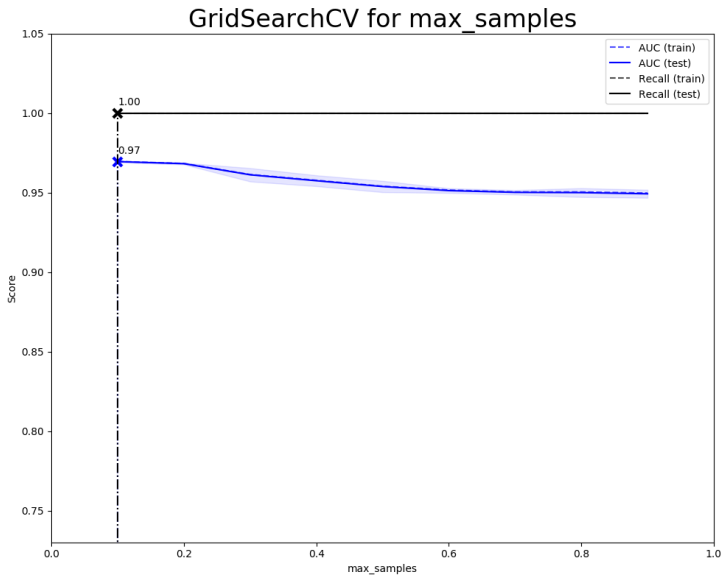
IF complexity on training is $O(num\_estimators \cdot sample\_size \cdot \log sample\_size)$.
IF complexity on testing is $O(num\_estimators \cdot n\_test \cdot \log sample\_size)$ [3].

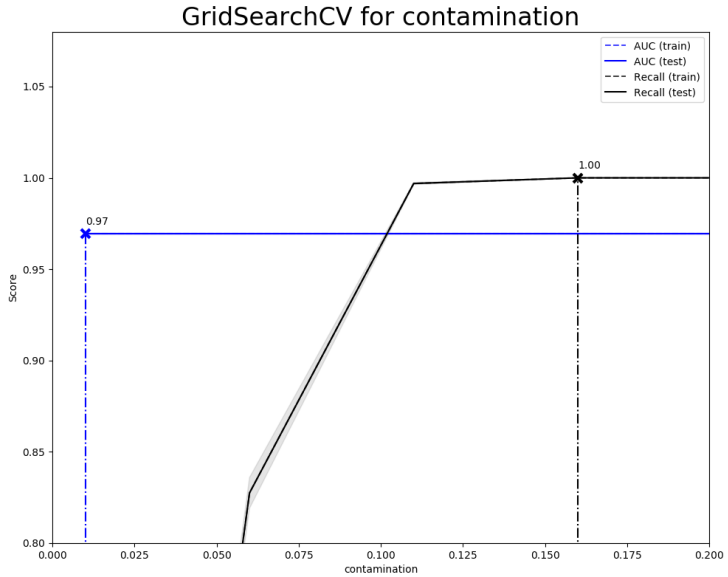
Overall, IF is very robust to parameterisation. One has to (mostly) configure:

- max_samples: number of samples to draw from X to train each base estimator
- n_estimators: number of base estimators in the ensemble
- contamination: the proportion of outliers in the data set (when using as supervised algorithm).

PARAMETERISATION CONT.







Robust results on different n_estimators and max_samples parameters. Contamination rate is key to control Recall and FP rate.

AML

Automated ML detection is not a simple *anomaly detection* problem and it is not really an *outlier detection* problem.

> Many patterns of transactions associated with money laundering differ little from legitimate transactions [4].

Outlier are often hidden in the unusual local behaviour of low-dimensional subspaces.

DREAMWORKS PICTURES

The core principle of discovering outliers is based on assumptions about the structure of the normal patterns in a given data set [1]. The choice of normal depends on the subject matter (banking type, clients type, markets performance, time of year, etc.).

Don't confuse extreme-value analysis and outlier analysis (e.g. $\{1,1,2,50,98,99,99\}$). Scenarios that monitor for unusually large transfer amount or unusually high frequency of trading/transferring often will only result in high FP rate.

PERFORMANCE ON SYNTHETIC DATA SET

Synthetic data set contains made-up transactions with some exhibiting known red flags.

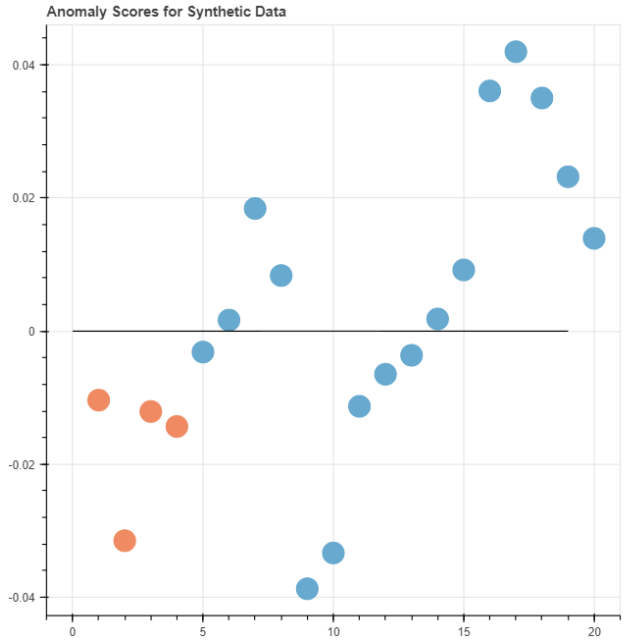
For more red flags see:

- https://www.dfs.ny.gov/
- http://www.fatf-gafi.org/
- http://www.jmlsg.org.uk/
- https://www.wolfsberg-principles.com/

The core idea is to use red flags as data models to transform trades/transactions data into lower dimensional sub-space. No need for many factors, as simple two-dimensional summaries by red flag/model can work well.

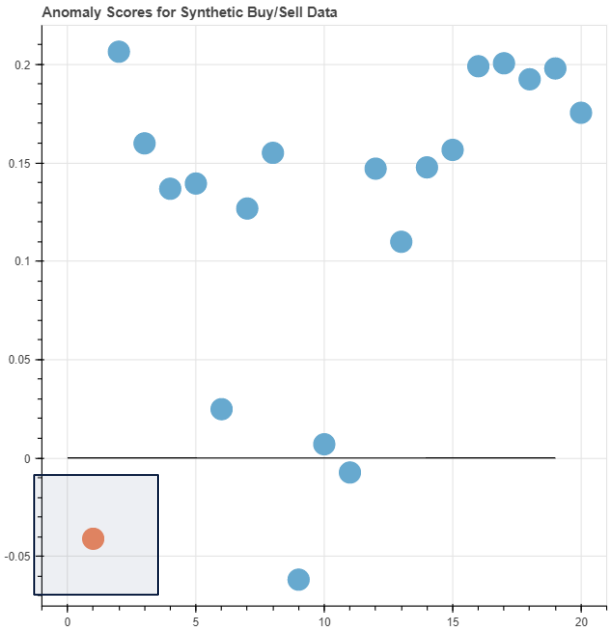
jupyter notebook can be shared on request. Contact me: www.codefying.com

All Synthetic Data

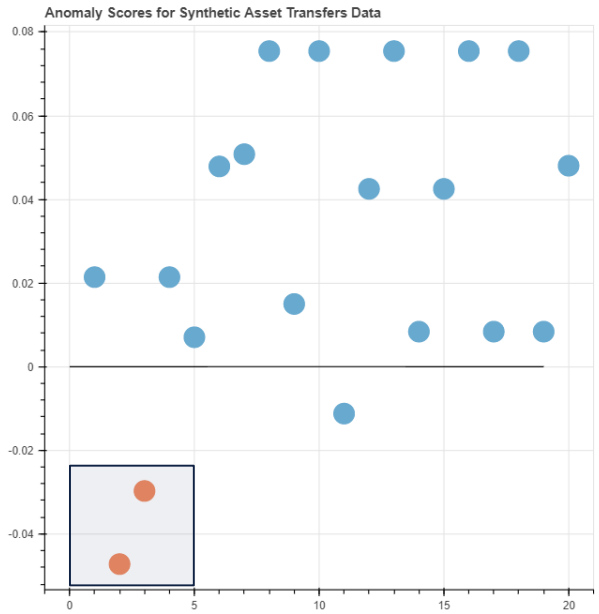


PERFORMANCE ON SYNTHETIC DATASET CONT.

Scenario 1



Scenario 2

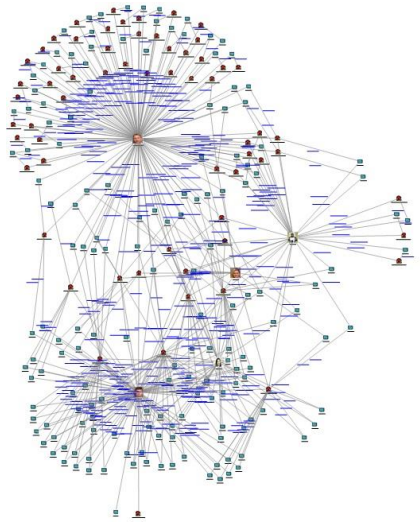


APPLICATION EXTENSION

Outliers can be masked by a full-dimensional analysis.

Another way to reduce dimensionality is to take a risk-based approach. This can mean to focus on high-risk client groups, certain client networks (e.g. entities that share the same beneficial owner, etc.).

For an IF-based analysis this would mean to introduce the network/connectivity factor, and analyse for red flags within a network, or one network vs. rest.



CONCLUSION

- Anomalous activity in AML is bank-type specific.
- Isolation Forest shows promise, as a fast and robust anomaly detection tool.
- In AML, successful automatic detection starts with asking the right questions about what is truly unusual and building a set of data models to mimic this: exploring low dimensional sub-spaces with red flag factors.

THANK YOU!

REFERENCES

[1] C.C. Aggarwal, Outlier Analysis, 2nd edition, Springer, 2017.

[2] A. Sudjianto et al., Statistical Methods for Fighting Financial Crimes, Technometrics, vol. 52, 2010.

[3] F.T. Liu, et al., Isolation Forest, Data Mining, 2008. ICDM'08, Eighth IEEE International Conference.

[4] U.S. Congress, Office of Technology Assessment, Information Technologies for the Control of Money Laundering, Sep. 1995, OTA-ITC-630.

[5] R.J. Bolton and J.D. Hand, Statistical Fraud Detection: A Review, Statistical Science, 17, 2002.