Outliers detection in complex networks via modularity

B. Kamiński, P. Prałat, F. Théberge, S. Zając

Decision Analysis and Support Unit
SGH Warsaw School of Economics, Poland

13-15.09.2022

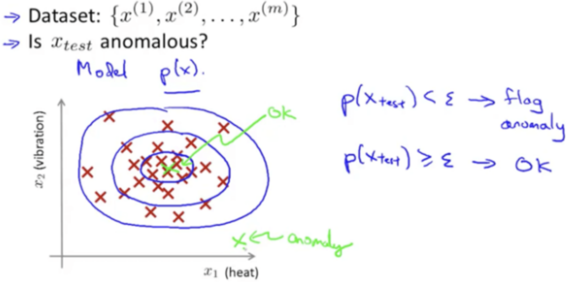
Outline of talk

1. Motivations
   - Anomaly detection
   - Anomaly detection in business
2. Community in complex networks
3. Synthetic networks generators
4. Realistic networks
5. Summary

Motivations

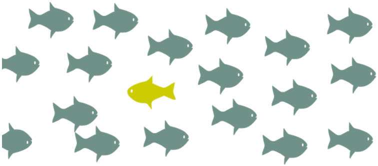
Anomaly detection

Anomaly detection is a technique used in data analysis for identifying **unexpected behaviour**, **outliers**, **rare events**, or **deviant objects**. Offline and online ML, DL, Quantum Computing methodology.



R. Foorthuis. On the Nature and Types of Anomalies: A Review of Deviations in Data. International Journal of Data Science and Analytics, 12(4) (2021)

Business Motivations



Anomaly in business

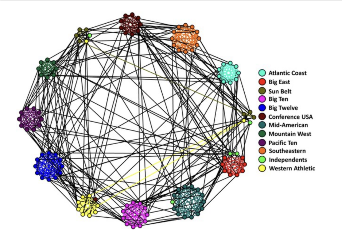
- Unusual Customer Behavioral Patterns
- Fraud detection
- Time series anomalies - stock price, climatology, epidemiology
- Monitoring - cardiac, machine condition, material quality, sales analysis

Our Case

Outliers detection in complex networks

Community in complex networks

Being able to identify communities in a network could help us to exploit it more effectively. Community structure plays an important role in understanding the properties of networks.



- social networks - groups by interest
- citation networks - related papers
- web communities - search engine, pages on related topics, fake news detection

A network has community structure if its set of nodes can be split into a number of subsets such that each subset is densely internally connected.

Even small number of nodes = a lot of partitions to consider.
Number of partitions? How to find good partitions?

First - historical approach

The set of nodes $C \in V$ forms a **strong community** if each node in C has more neighbours in C than outside of C: $\deg^{\text{int}}(v) > \deg^{\text{ext}}(v)$.
C forms a **weak community** if the avg degree inside the community C (over all nodes in C) is larger than the corresponding avg number of neighbours outside of C.

In this context, **an outlier** could be defined as a node that does not have majority of its neighbours in any of the communities.

Community detection algorithms, Modularity

Approach using definition of outliers using *strong community* approach is too strict as it typically would lead to too many nodes identified as outliers.

Modularity

Community detection can be based on modularity function.
Modularity for graphs is based on the comparison between the actual density of edges inside a community and the density one would expect to have if the nodes of graph were attached at random (Chung-Lu null-model).
For a given partition $\mathsf{A} = \{A_1, A_2, \ldots, A_\ell\}$ of V the modularity function is as follows:

$$q_o(\mathsf{A}) = \sum_{i=1}^{\ell} \frac{e(A_i)}{|E|} - \sum_{i=1}^{\ell} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)} \right)^2$$

where e(A) is the number of edges within set A, e(A,B) is the number of edges between set A and B, and vol(A) is the sum of degrees of nodes in A.

Problems and business case

There are several efficient approaches for community detection that use modularity, the most popular are Louvain, Leiden, and ECG.

Each node is forced to be a member of some community, but in practice some nodes can be outliers (do not fit well into any community) – how to find them?
Such non-fitting nodes can be outliers (do not fit anywhere) or fit several communities (overlapping communities approach)

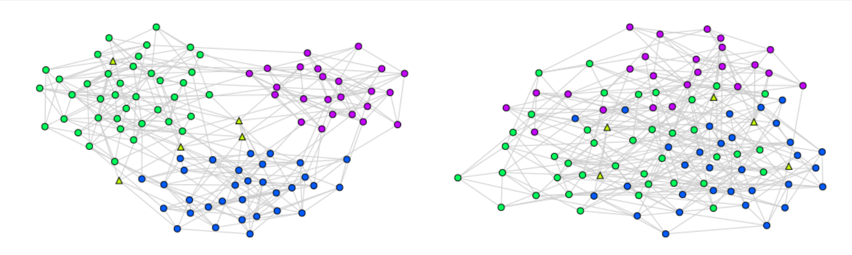
Business case

- Tesla fans and BMW fans - but what about fans of all cars? (outlier)
- Barcelona fans Real and Madrid fans - but what about fans of Lewandowski?

Synthetic networks for Experiments

ABCD Random Graph Model with community structure

The **A**rtificial **B**enchmark for **C**ommunity **D**etection graph is a random graph model with community structure and power-law distribution for both degrees and community sizes. It has been recently augmented to allow for generation of outlier nodes (ABCD+o).



ABCD+o graphs with ($\xi = 0.2$, left) and ($\xi = 0.4$, right). The number of outliers is $s = 5$.

See also: ABCD graph generator in Julia programming language - `https://github.com/bkamins/ABCDGraphGenerator.jl`
B. Kamiński, P. Prałat, F. Théberge: „*Mining Complex Networks*”, CRC Press (2022) or *Outliers in the ABCD Random Graph Model with Community Structure (ABCD+o)*.

Real networks for Experiments

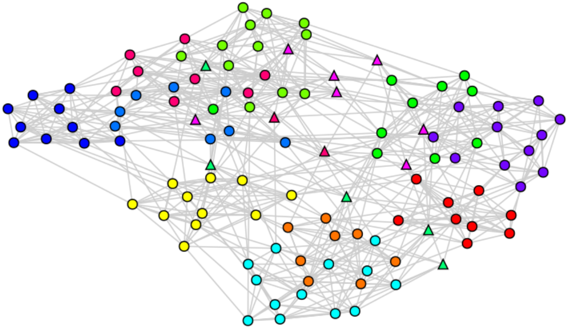


Fig. 3: The College Football Graph; outliers are displayed with triangular shape.

First attempt

How do adjust modularity to take into account possible outliers?
We identify nodes that have a distribution of edges approximately evenly spread among all communities and consider them to be outliers not assigned to any community.

For a given partition $\mathsf{A} = \{A_1, A_2, \ldots, A_\ell, O\}$ of V, where O is the set of outliers, the *modularity function* is adjusted as follows:

$$
\begin{aligned}
q_o(\mathsf{A}) \;&=\; \sum_{A_i \in \mathsf{A}} \frac{e(A_i)}{|E|} - \sum_{A_i \in \mathsf{A}} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)} \right)^2 \\
&\quad -\lambda \left(\frac{e(O)}{|E|} - \left(\frac{\mathrm{vol}(O)}{\mathrm{vol}(V)} \right)^2 \right).
\end{aligned}
$$

where $\lambda \in \mathsf{R}^+$ is a regularisation parameter.

Summary

Feature (and present) work

1. New definition of the modularity function with outliers.
2. New scalable optimization algorithm for outlier detection.
3. Synthetic networks such as ABCD-o and null models. Investigating experimentally and theoretically their properties.
4. Analyzing real networks and business applications.

If you are interested in this topic, have some suggestions/remarks, then please contact our group.

Thanks for Your Attention!
sebastian.zajac@sgh.waw.pl

This work has been supported by the Polish National Agency for Academic Exchange under the Strategic Partnerships programme, grant number BPI/PST/2021/1/00069/U/00001.