Outlier detection with community structure on graphs

B. Kaminski, P. Pralat, F. Théberge, S. Zajac

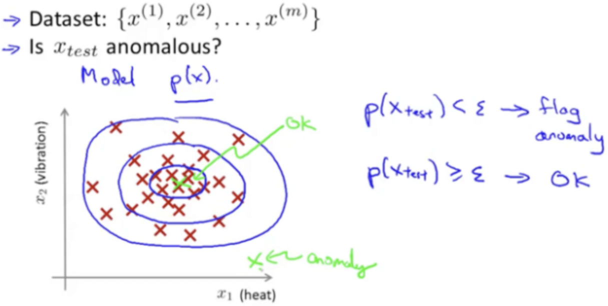
WAW 2023, The Fields Institute
25.05.2023

This research was supported by the Polish National Agency for Academic Exchange under the Strategic Partnerships programme, grant number BPI/PST/2021/1/00069/U/00001.
Material preparation, slide formatting & layout supported by:
Marta Jaron-Chruslinska & Boguslawa Wasik-Szczygiel.

Motivations

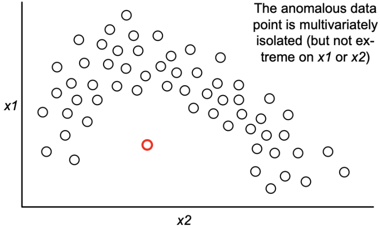
Anomaly detection

Anomaly detection is a technique used in data analysis for identifying **unexpected behaviour**, **outliers**, **rare events**, or **deviant objects**.

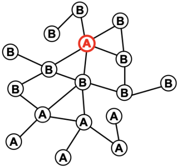


R. Foorthuis. On the Nature and Types of Anomalies: A Review of Deviations in Data. International Journal of Data Science and Analytics, 12(4) (2021)

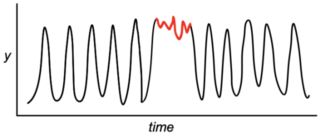
Motivations



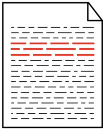
The anomalous data point is multivariately isolated (but not extreme on *x1* or *x2*)



The anomalous vertex has a different class label than its adjacent vertices.



The anomalous time interval deviates from the cyclical pattern.



The anomalous text section is comprised of unusually long words.

Science and Business Motivations

Cybersecurity: attacks, malware, malicious apps/URLs, biometric spoofing

Social Network and Web Security: false/malicious accounts, false/hate/toxic information

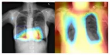
Video Surveillance: criminal activities, road accidents, violence, etc.

fighting *road accident*

shooting *shoplifting*

Finance: credit card/insurance frauds, market manipulation, money laundering, etc.

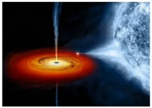
Healthcare: lesions, tumours, events in IoT/ICU monitoring, etc.



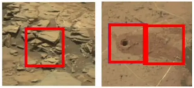
Industrial Inspection: Defects, micro-cracks

Drug Discovery: rare active substances

Astronomy: Anomalous events



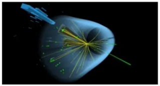
Rover-Based Space Exploration: unknown textures



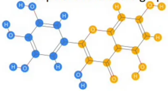
Bedrock (Sol 1032)

Drill hole and tailings (Sol 1496)

High-Energy Physics: Higgs boson particles



Material Science: exceptional molecule graphs



Methodology

Outlier research has a long history (Bernoulli - 1777) and traditionally focused on techniques for rejecting or accommodating the extreme cases that hamper statistical inference.

- Online and offline for tabular, unstructured and graph data
- Supervised, unsupervised, semi-supervised for ML and DL
- CPU, GPU, Quantum computers

Graph-based Anomaly Detection [Surveys]

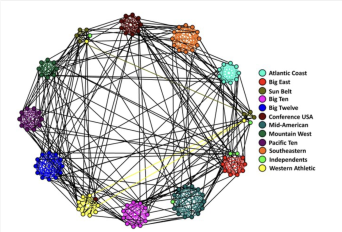
- A comprehensive Survey on Graph Anomaly Detection with Deep Learning. IEEE TKDE Sep 2021
- Graph-based Anomaly Detection and Description: A Survey. DAMI, May 2015
- Fraud Detection through Graph-Based User Behavior Modeling. ACM CCS 2015
- Anomaly, Event, and Fraud Detection in Large Graph Datasets, ACM WSDM 2013

Our Case

Outliers detection in networks for static and simple graphs data

Community in complex networks

Identifying communities in a network could help us to exploit it more effectively. Community structure plays an essential 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

Community in complex network

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

Even a small number of nodes = a lot of partitions to consider.

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 C.

$$\sum_{v \in C} \deg^{\text{int}}(v) > \sum_{v \in C} \deg^{\text{ext}}(v)$$

Community detection algorithms, Modularity

In this context, **an outlier** could be defined as a node that does not have most of its neighbours in any of the communities. Using a *strong community* approach typically would lead to **too many nodes** identified as outliers. No dependence on community size

Modularity

Community detection can be based on a modularity function. **Modularity** for graphs is based on the comparing the actual density of edges inside a community and the density one would expect to have if the graph nodes were attached at random (Chung-Lu null-model).

Modularity

Standard modularity definition

$G = (E, V)$, for a given partition $A = \{A_1, \ldots, A_\ell\}$ of V the modularity function is defined as:

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

where e(A) is the **number of edges within set A**, $\mathrm{vol}(A)$ is the **sum of degrees of nodes** in A, and $e(V) = |E|$.

edge contribution – degree tax

The $q(A)$ function is maximized over the set of all partitions of V to find optimal split of the graph into communities.

Resolution limit

Optimization of $q(A)$ is prone to the resolution limit. Optimizing modularity function in large networks cannot find small communities, even if they are well defined.

- To overcome this problem, we can use $\lambda > 1$ then we penalize large communities more.
- $\lambda \to \infty$ = each node as a community.

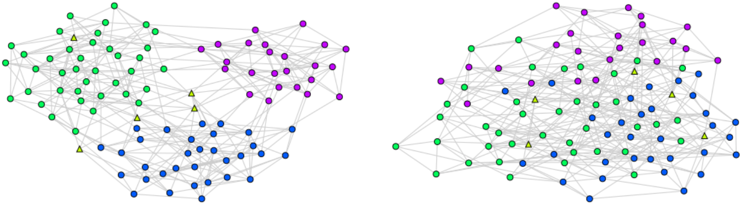
$$q(A) = \sum_{i=1}^{\ell} \frac{e(A_i)}{|E|} - \lambda \sum_{i=1}^{\ell} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)} \right)^2$$

Experiment with random graph

Preparing graph $G(V, E)$ with $\xi = 0.2$, 1000 nodes and 8773 edges with 25 outliers. Apply Leiden algorithm with different λ values:

ABCD+o

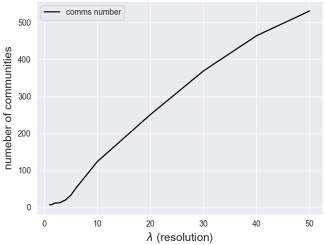
The Artificial Benchmark for the Community Detection graph is a random model with community structure and power-law distribution for degrees and community sizes. It has been recently augmented to allow for the generation of outlier nodes (ABCD+o).

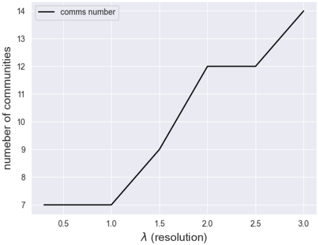


ABCD+o graphs with ($\xi = 0.2$, left) and ($\xi = 0.4$, right)

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)*.

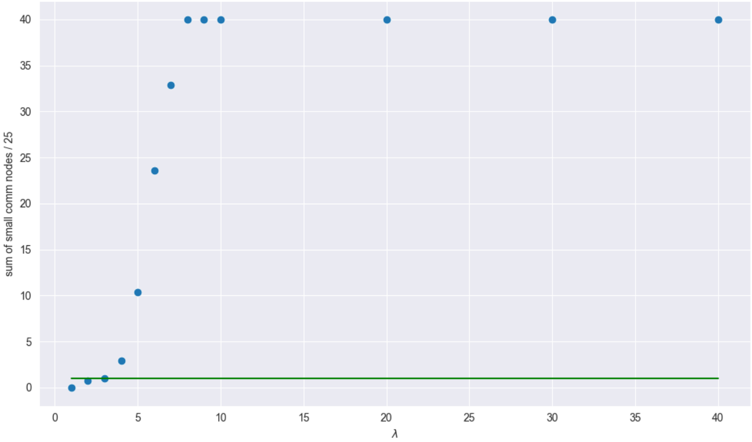
Lambda resolution - experiment

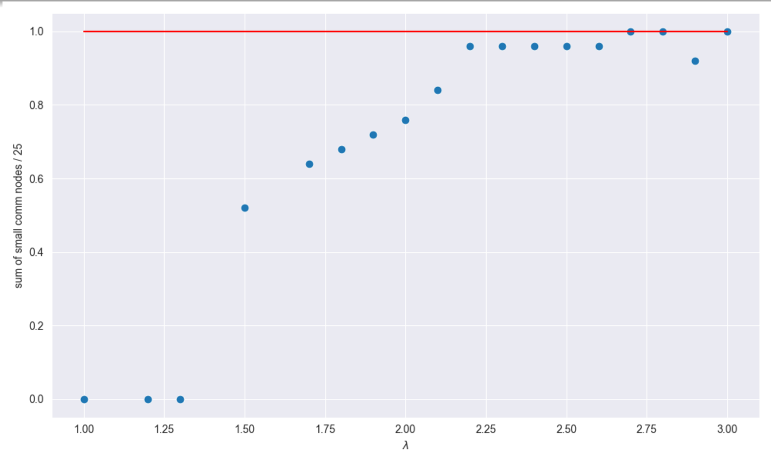




for $\lambda = 1$ - 7 communities ('1':2:104, '2':3:186, '3':3:183, '4':5:109, '5':1:151, '6':5:108, '7':6:159)
for $\lambda = 2$ - 12 communities ('1':0:102, '2':1:184, '3':0:180, '4':3:107, '5':0:150, '6':2:105, '7':0:153, '8':7:7, '9':6:6, '10':4:4, '11':1:1, '12':1:1)

Lambda resolution - experiment





our task

how can it be modified to find small communities without changing typical, huge communities?

Outliers score - Modularity modification

We propose modularity modification. In general case:

$$q(A) = \sum_{A_i \in \mathrm{A}} \frac{e(A_i) + \beta * [|A_i| \leq \delta] * \mathrm{vol}(A_i)/2}{|E|} - \lambda \sum_{A_i \in \mathrm{A}} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)} \right)^2$$

The $[\ell]$ notation evaluates to 1 is ℓ is true and to 0 otherwise. In simple case when $\delta = 1$ we try separate 1 node community. For a given node v we can compute β^* that satisfy:

$$\beta^* = 2\frac{e_v}{\deg(v)} - 2\lambda \frac{\mathrm{vol}(A_i) - \deg(v)}{\mathrm{vol}(V)}$$

So outliers have a low fraction of within-community edges, while at the same time, moving them to another community does not improve this situation

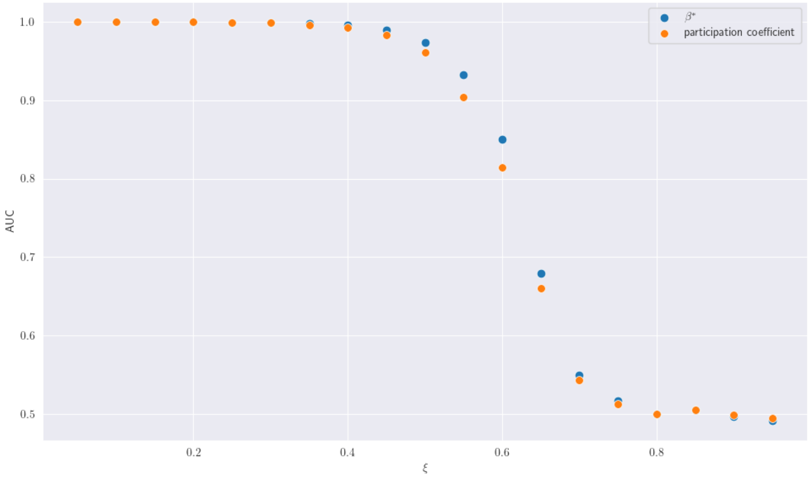
We use ABCD+o generator for graphs:

- 20 000 nodes,
- 100 outliers,
- degree distribution: min 6, max 2000,
- community size distribution: min 1000, max 3000,
- 64 times test for $\xi \in [0.05, 0.95]$ with step 0.05.

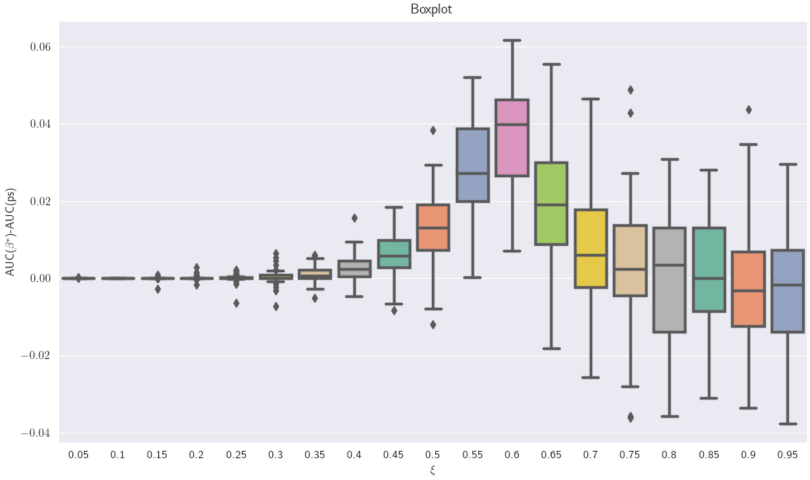
Let's defined the **participation coefficient** of a node v as:

$$p(v) = 1 - \sum_{i=1}^{\ell} \left(\frac{\deg_{A_i}(v)}{\deg(v)} \right)^2$$

ABCD+o results



ABCD+o results



Thanks for Your Attention!

sebastian.zajac@sgh.waw.pl