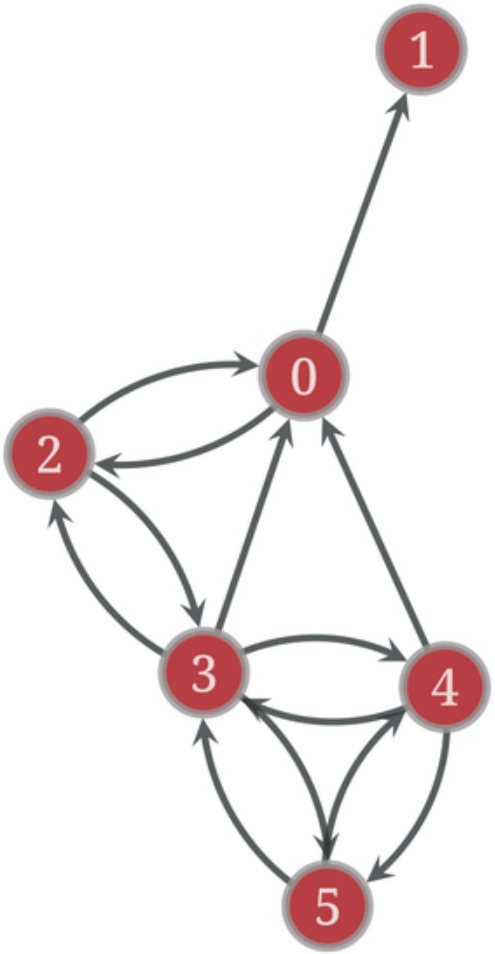
Calculation of Graph Isomorphisms in the Context of Processing of Big Structured Data

Igor Wojnicki, Aleksander Suchorab, Andrzej Bielecki, Marzena Bielecka

2025-04-03

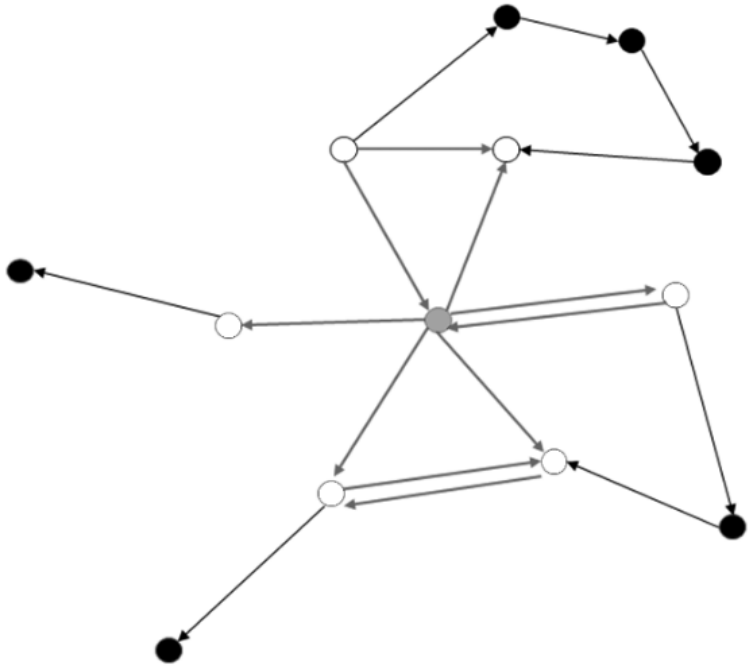
Problem Statement



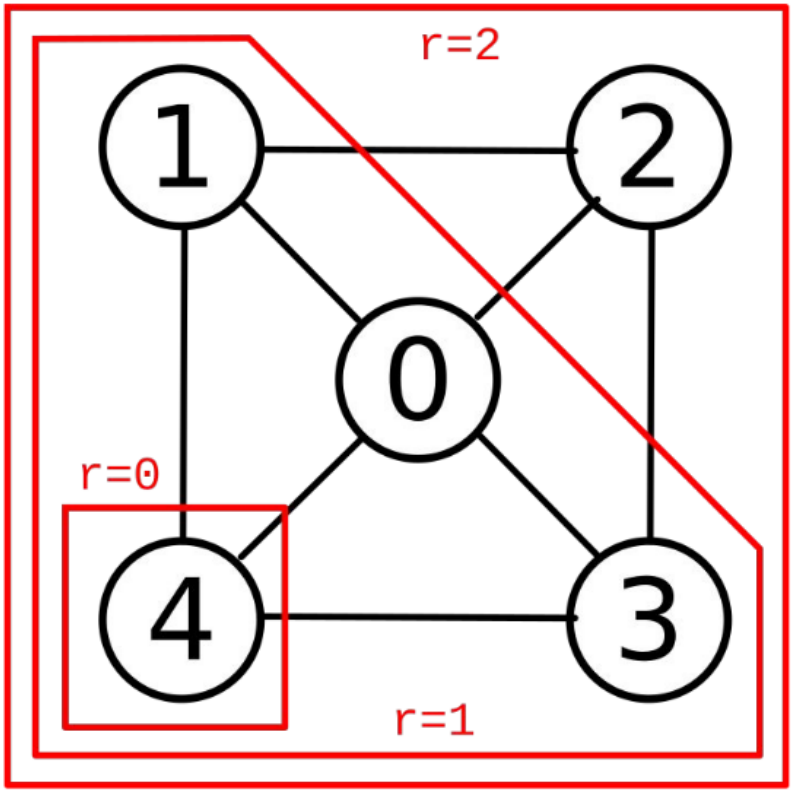
- How much information is in a graph?
  - Is it suitable for Machine Learning?
  - Will it work with particular Graph Embeddings?
  - Will it work with Graph Neural Networks.
- *Node-ball isomorphism*
  - Is it different than *graph entropy*?

Structural Information

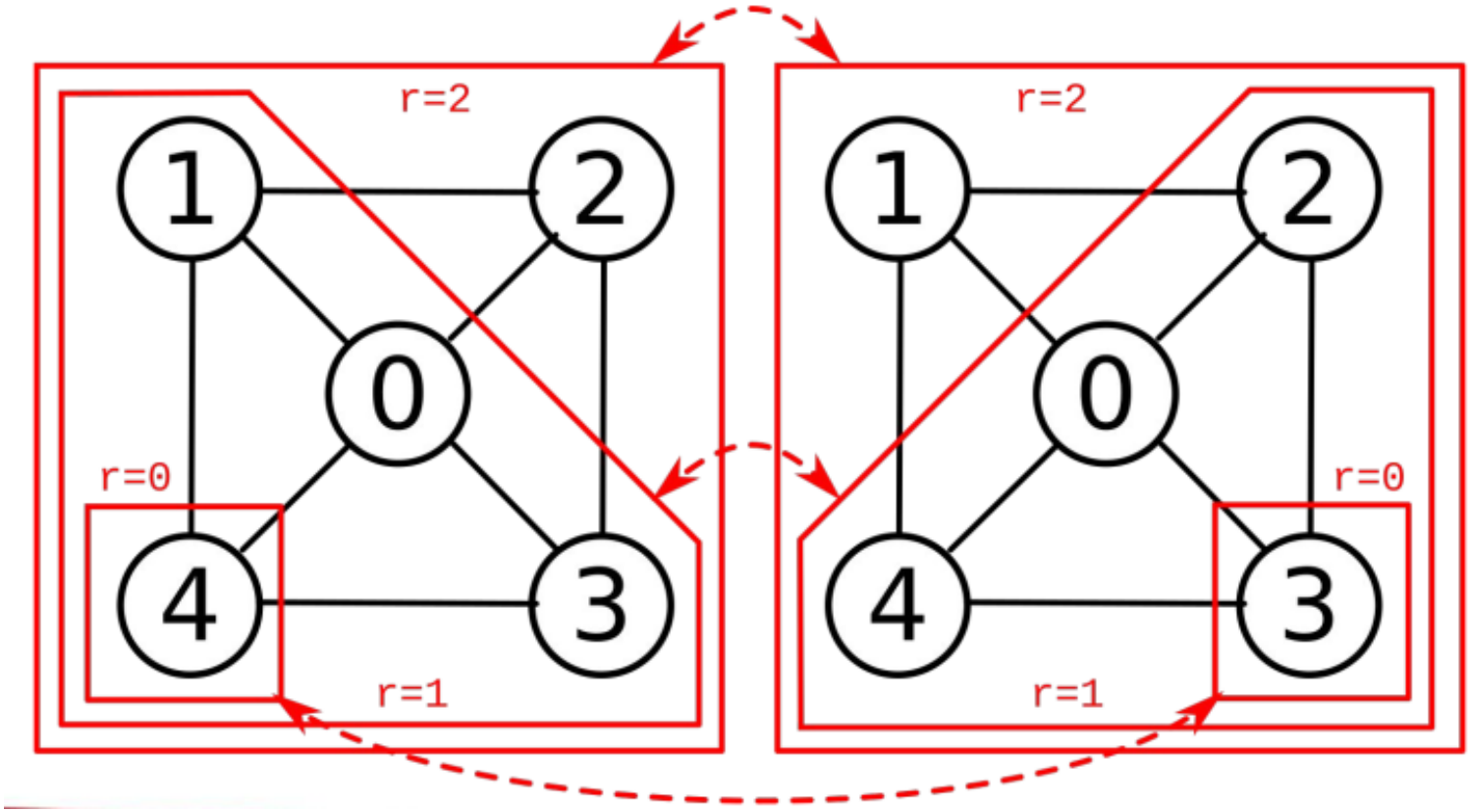
- What is a node-ball?
- Two nodes are not distingusishable
  - if all node-balls of the same radii are isomorphic.



Node-Balls



Node-Ball Isomorphism



- Processing.
  - Multiple graph isomorphisms must be calculated.
  - The isomorphims must be anchored.

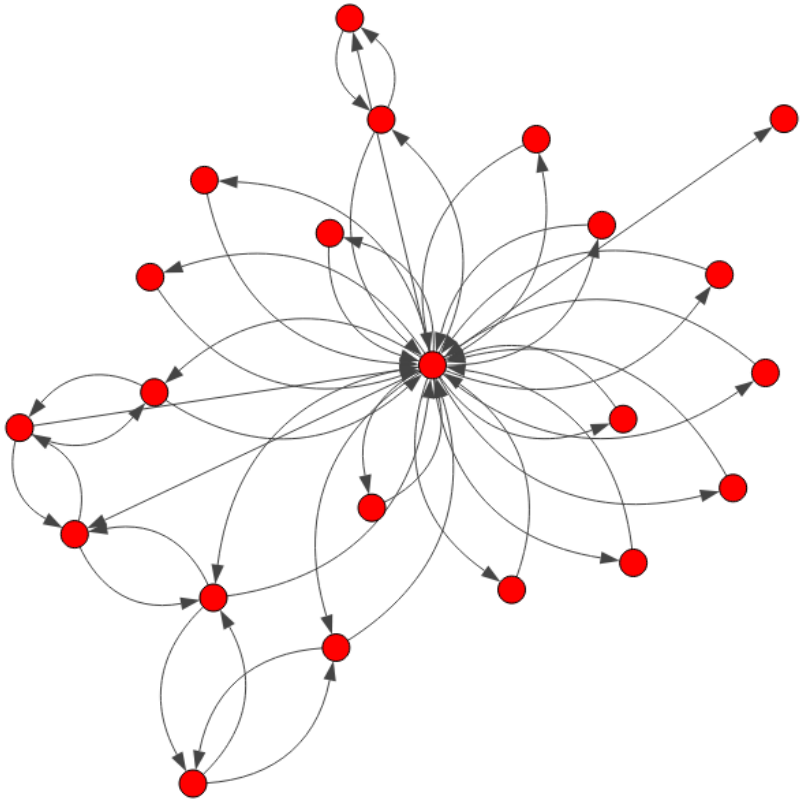
How much information is there?

- Equivalence classes.
  - All indistingusishable nodes belong to the same equivalence class.
- If there are K classes, the Hellerman's factor is:

$$H^{node} = -n \sum_{k=1}^{K} \frac{n_k}{n} log_2(\frac{n_k}{n})$$

- n – number of nodes, n_k – numner of elements in the k-th class.
- Same as *graph entropy*.

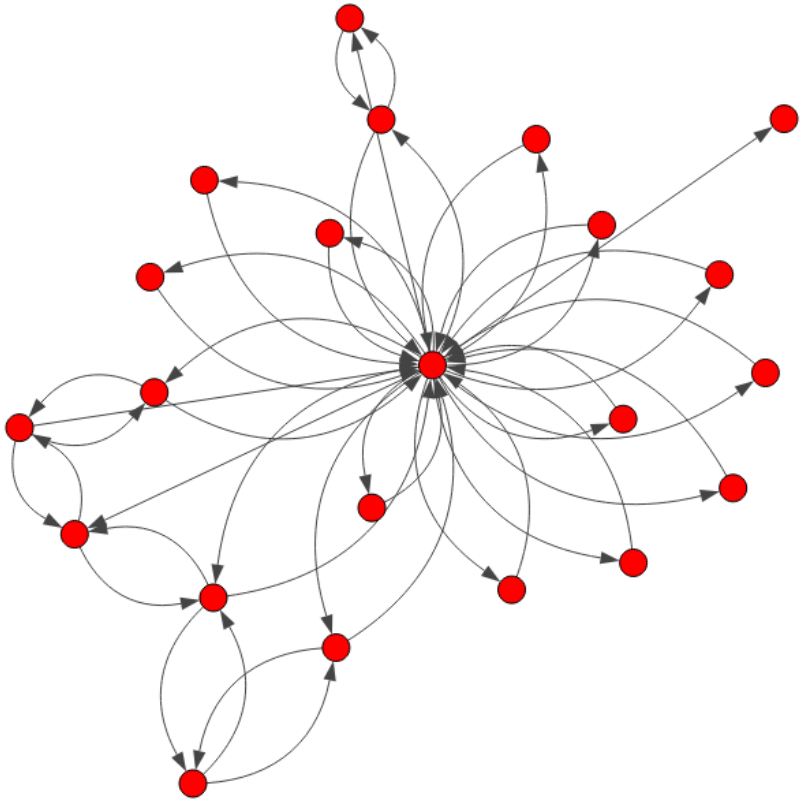
Scale Problem, Find All Isomorphisms



- How many isomporphisms there are?
- 479 001 600

AGH

Scale Problem, Find All Isomorphisms



- How many isomporphisms there are?
- 479 001 600

Technological Evolution

- PyKeen
  `https://github.com/pykeen/pykeen`
- networkx `https://networkx.org/`
- igraph `https://igraph.org/`
- graph-tool
  `https://graph-tool.skewed.de/`
- C++/MPI, Cyfronet Ares

Algorithms

- Non-strict
- Half-strict, polynomial
- Strict, exponential.

Results, reverse square root of H

Dataset	Węzły	Krawędzie	No labels		
			non-strict	half-strict	strict
Countries	210	964	0.290711	0.274022	0.274022
CodExSmall	2034	36543	0	0	0
PharmKG8k	7247	485787	0.024440	0.024440	0.024440
OpenEA	14279	37502	0.143415	0.143246	0.143247
FB15k237	14496	310074	0.087427	0.077148	0.077148
FB15k	14933	592185	0.093249	0.093197	0.093197
CodExMedium	17050	206205	0.016348	0.016348	0.016348
DBpedia50	19938	30447	0.383260		0.382790
WD50KT	40103	232342	0.194599	0.193509	0.193340
WN18RR	40533	92566	0.181255		0.179803
WN18	40917	151414	0.186133		0.184675
AristoV4	41986	279410	0.309766	0.307321	0.307202
Hetionet	45158	2250197	0.122941	0.122913	0.122913
CodExLarge	77940	612404	0.062458		0.061631
DB100K	99378	697177	0.141773		

Summary

- Tests on actual data sets confirm suitability.
- Multiple versions of the algorithm have been tested, including approximations.
- We didn't get old in the process, 46 879 CPU hrs.

Yet to do:

- investigation of partial information content for $r < r_{max}$,
- edge related information content: *edge-balls*.