Discovering alignment relations with Graph Convolutional Networks

A case study in pharmacogenomics

Pierre Monnin, Chedy Raïssi, Amedeo Napoli, Adrien Coulet

✉ adrien.coulet@inria.fr

June 29, 2022

Loria Laboratoire lorrain de recherche en informatique et ses applications

Inria

Inserm

ANR — Projet financé par l'ANR – Project funded by the ANR

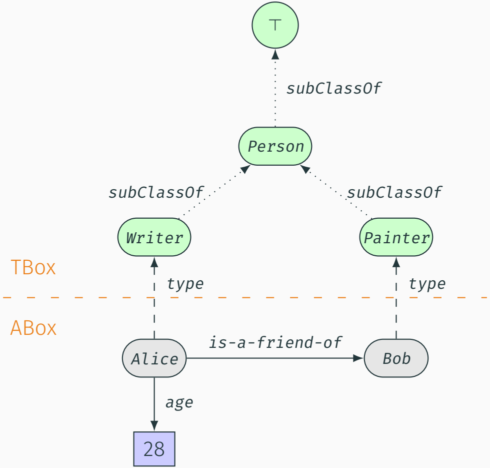
Context: Knowledge graphs in the Web of data

- Directed and labeled multigraphs
- Nodes

  Individuals   Classes   Literals
- Edges
  - Labeled by a predicate
  - Defined by triples

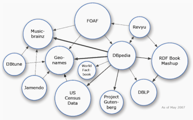
$$\langle subject, predicate, object \rangle$$

- Semantic Web standards
  RDF, URI, RDFS, OWL, SPARQL, ...
  (Berners-Lee et al. 2001)

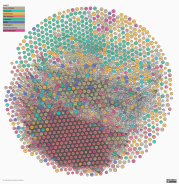


Motivation: From a general assessment about the Web of data...

Increasing size and number of available knowledge graphs
→ Interest in their conjoint use



(a) LOD Cloud in 2007



(b) LOD cloud in 2020

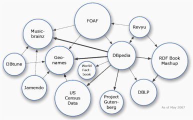
Motivation: From a general assessment about the Web of data...

Increasing size and number of available knowledge graphs

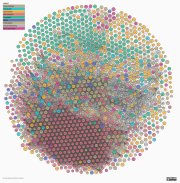
→ **Interest in their conjoint use**

- Concurrent publication and edition → possible overlap
- Heterogeneity issues: vocabularies, granularities, ...

→ **Matching similar units within and across KGs**



(a) LOD Cloud in 2007

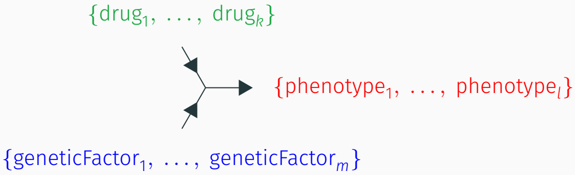


(b) LOD cloud in 2020

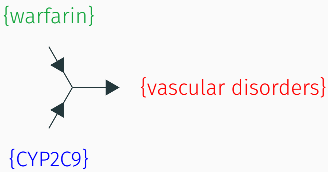
from `lod-cloud.net` –

Pharmacogenomics (PGx) studies the influence of gene variants on drug responses

Abstract PGx relationship



Example: CYP2C9 and warfarin



… to a specific application: pharmacogenomics (PGx)

- 2 sources of PGx knowledge

- Interest of matching
  - PGx knowledge useful in precision medicine
  - State-of-the-art knowledge may lack validation

  → Align sources to obtain a consolidated view of the PGx knowledge

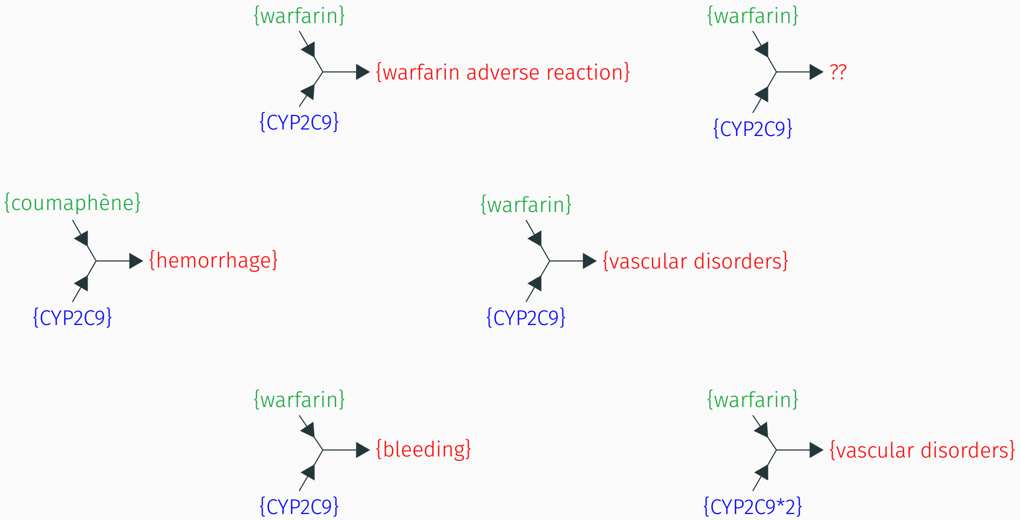
- Such a view can then be mined
  - Explain Adverse Drug Reactions
  - Predict side effects or pharmacogenes

Specialized databases

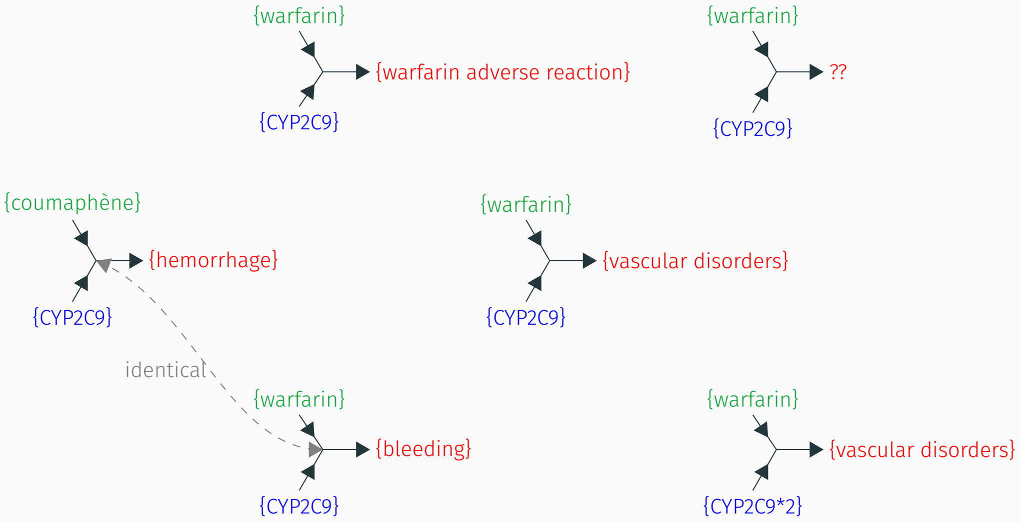
(*e.g.*, PharmGKB)

Biomedical litterature

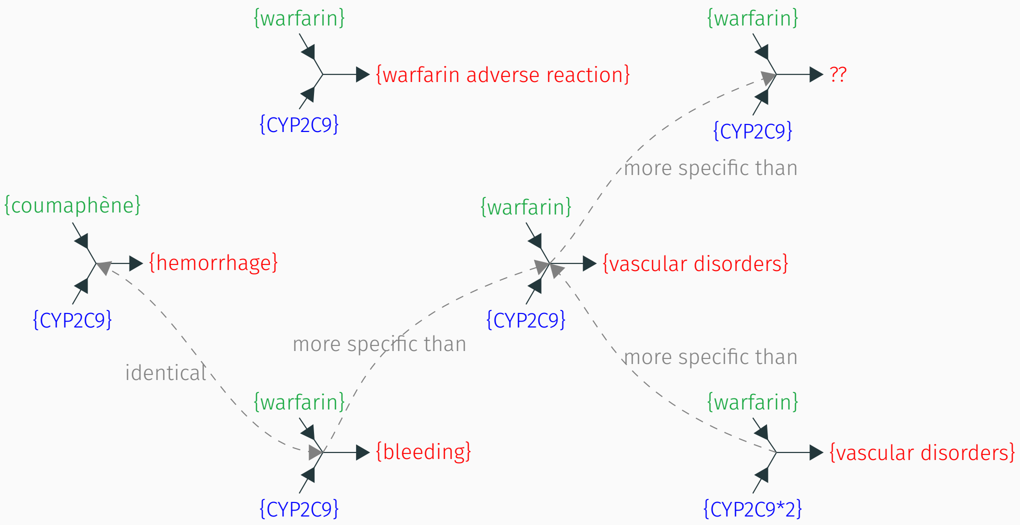
Example of expected matching results between PGx relationships



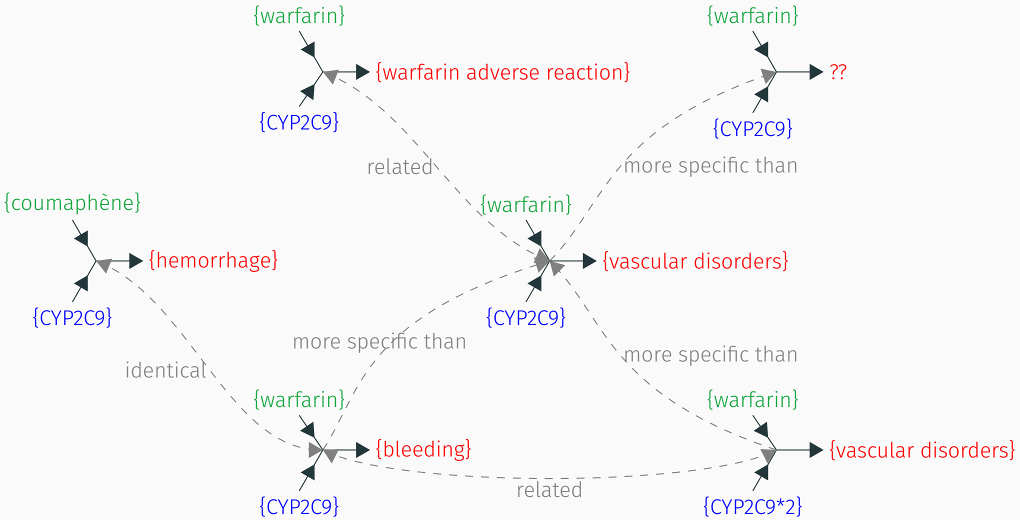
Example of expected matching results between PGx relationships



Example of expected matching results between PGx relationships



Example of expected matching results between PGx relationships



Preliminary work

PGxLOD: a knowledge graph for pharmacogenomics

PGxLOD

PGxO class	Number of instances
PharmacogenomicRelationship	50,435
↳ *From PharmGKB (structured data)*	3,650
↳ *From PharmGKB (text mining)*	10,240
↳ *From the literature*	36,535

88M triples

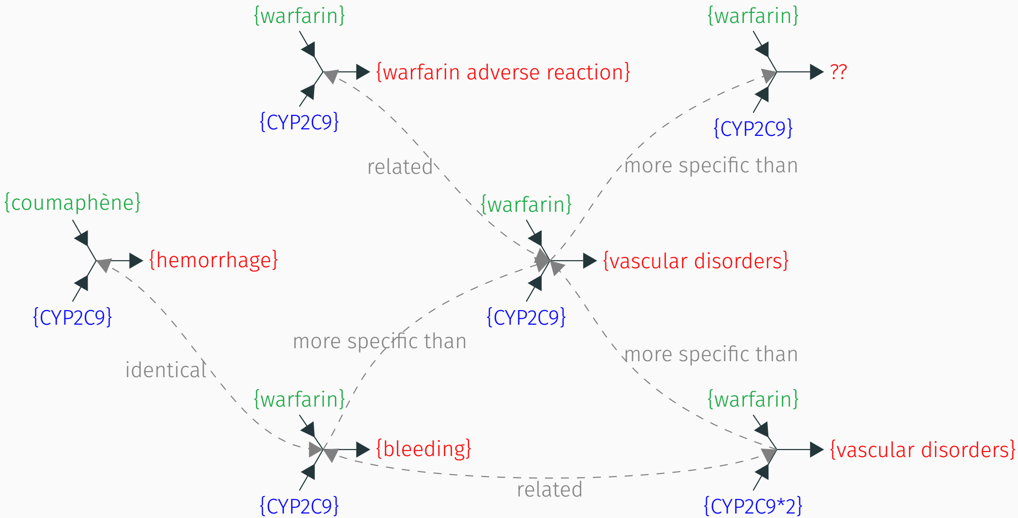
11M nodes (w/o literals)

416 predicates

pgxo.loria.fr pgxlod.loria.fr practikpharma/pgxo

(Monnin, Jonquet, et al., *NETTAB*, 2017) (Monnin, Legrand, et al., *BMC Bioinformatics*, 2019)

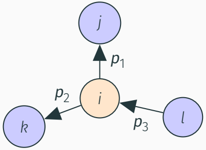
Example of expected matching results between PGx relationships



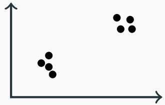
Match nodes with a structure-based approach

Discovering alignment relations with Graph Convolutional Networks

Graph embedding: definition & intuition



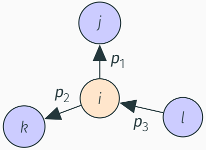
Graph structures
(*e.g.*, nodes)



d-dimensional space
preserving graph properties

(Cai et al. 2018; Chami et al. 2020; Ji et al. 2020; Nickel et al. 2016; Q. Wang et al. 2017)

Graph embedding: definition & intuition



Graph structures
(*e.g.*, nodes)

d-dimensional space
preserving graph properties

(Cai et al. 2018; Chami et al. 2020; Ji et al. 2020; Nickel et al. 2016; Q. Wang et al. 2017)

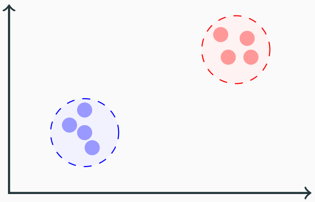
Intuition: Continuous aspect → enable flexibility (Guha 2015)

May capture fuzzier similarities, deal with missing direct mappings, ...

Graph embedding for matching nodes

Possible tasks

Link prediction (*e.g.*, ***`owl:sameAs`***), node classification, node clustering



Learning task

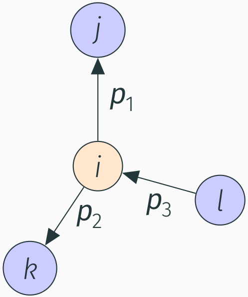
Learn node embeddings such that

similar nodes $\leftrightarrow$ low distance b/w their embeddings

Learning node embeddings with Graph Convolutional Networks (GCNs)

- "Message-passing framework" of multiple layers (Kipf et al. 2017; Schlichtkrull et al. 2018)
- Learns embeddings of a node *i* w.r.t. its neighbors

$$h_i^{(l+1)} = \sigma\left(\underbrace{\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)}}_{\text{Neighborhood}} + \underbrace{W_0^{(l)} h_i^{(l)}}_{\text{Self-connection}} \right)$$



→ **Well-adapted to a structure-based matching**
(Pang et al. 2019; Z. Wang et al. 2018)

Learning node embeddings with GCNs and the Soft Nearest Neighbor (SNN) loss

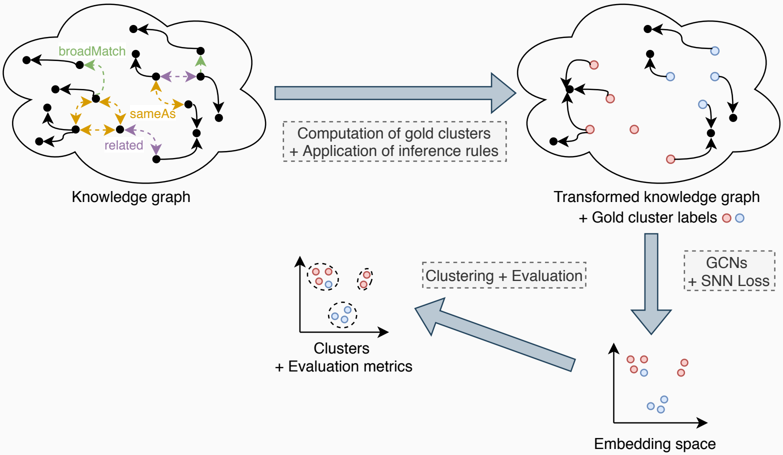
$$\min \mathcal{L}_{\mathrm{SNN}}(N, Y, T, h) = -\frac{1}{|N|} \sum_{i \in N} \log \left(\frac{\sum\limits_{\substack{j \in N \\ j \neq i \\ Y_i = Y_j}} e^{-\frac{||h_i - h_j||^2}{T}}}{\sum\limits_{\substack{k \in N \\ k \neq i}} e^{-\frac{||h_i - h_k||^2}{T}}} \right)$$

- h: node embeddings
- T: temperature
- N: nodes involved in clusters
- Y: cluster labels of nodes in N

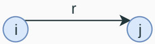
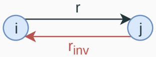
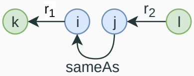
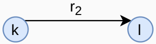
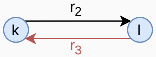
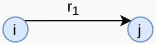
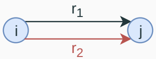
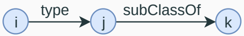
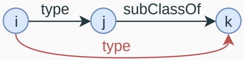
$\rightarrow$ Minimize intra-cluster distances and maximize inter-cluster distances

(Frosst et al. 2019)

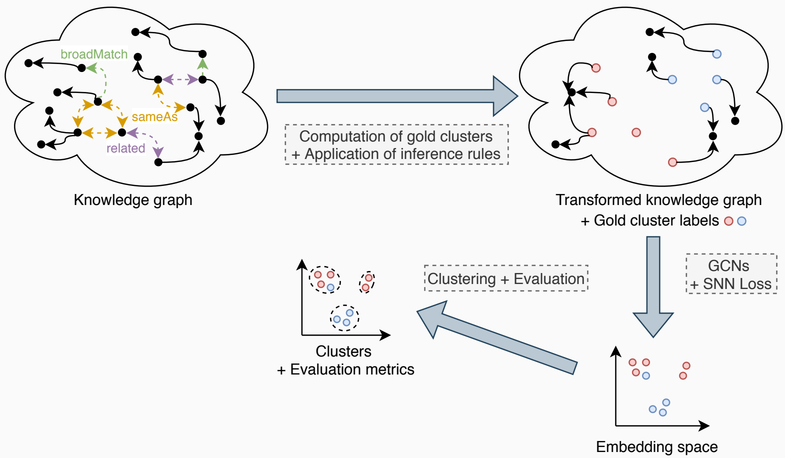
Approach outline



Applying inference rules associated with domain knowledge

Graph	Before	After
$\mathcal{G}_0$		
$\mathcal{G}_1$		
$\mathcal{G}_2$	$r_3 \equiv r_2^{-1}$ 	
$\mathcal{G}_3$	$r_1 \sqsubseteq r_2$ 	
$\mathcal{G}_4$		
$\mathcal{G}_5$	All transformations from $\mathcal{G}_1$ to $\mathcal{G}_4$	

Approach outline



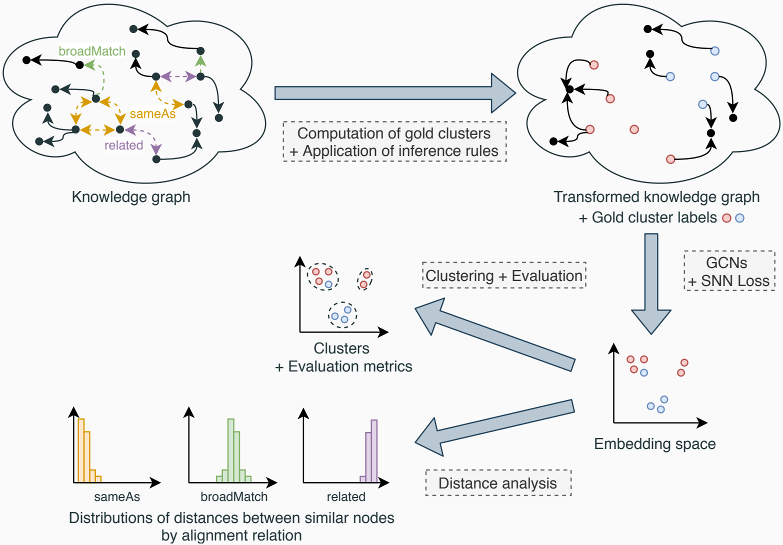
Clustering performance

Several clustering algorithms were evaluated (Ward, Single, OPTICS)
we only report results with Single (best results)
Different "gold clustering" were used as alternative gold standards
we only report results with $\mathcal{C}_0$ gold clustering (a mix of different alignement relations)

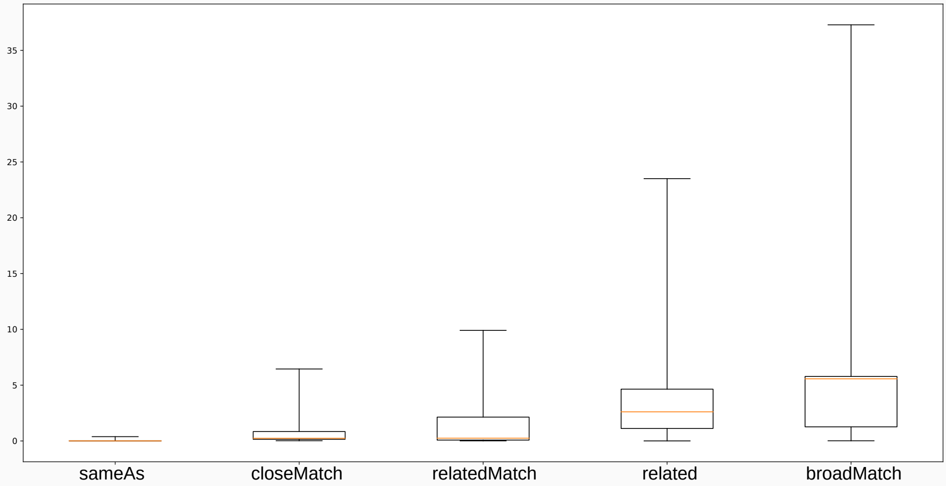
Gold clustering $\mathcal{C}_0$ on all graphs

Graph	Performance
$\mathcal{G}_0$ (no inference rules)	Baseline
$\mathcal{G}_1$ (*sameAs* contraction)	Improvements
$\mathcal{G}_2$ (inverses and symmetry of predicates)	Deterioration
$\mathcal{G}_3$ (hierarchy of predicates)	Improvements
$\mathcal{G}_4$ (hierarchy of classes)	Consistent deterioration
$\mathcal{G}_5$ (all inference rules)	Improvements – Best results

Approach outline



Distance analysis: rediscovery of alignment relations?



$\mathcal{G}_0$ – Fold 1

Conclusion & Perspectives

Conclusion & Perspectives

- Difficult task: uneven cluster sizes, different alignment relations
- **In difficult settings, domain knowledge improves performance**
  → Interest of considering domain knowledge within embedding approaches
- **Distances coherent w.r.t. "strength" of relatedness**
  → Emergence of semantics in the embedding space

pmonnin/gcn-matching

(Monnin, Raïssi, et al., *DL4KG@ESWC*)
(Monnin, Raïssi, et al., *Semantic Web Journal*)

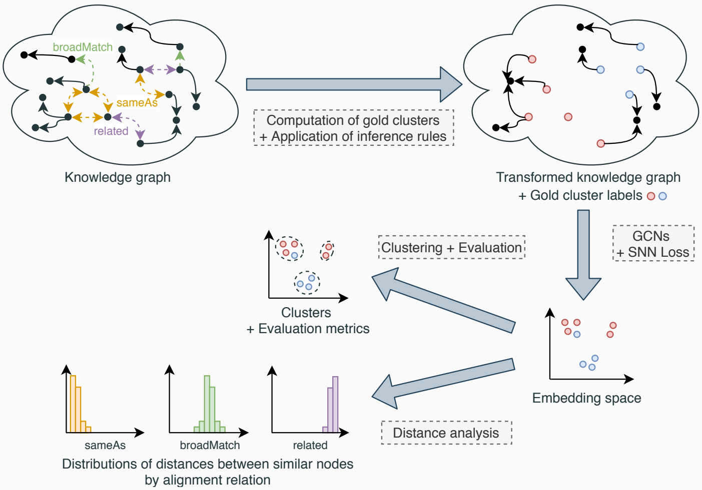
Conclusion & Perspectives

- Difficult task: uneven cluster sizes, different alignment relations
- **In difficult settings, domain knowledge improves performance**
  → Interest of considering domain knowledge within embedding approaches
- **Distances coherent w.r.t. “strength” of relatedness**
  → Emergence of semantics in the embedding space
- To confirm with other graph embedding techniques / other tasks
- Where to consider domain knowledge?
- What other semantics could emerge in the embedding space?

pmonnin/gcn-matching

(Monnin, Raïssi, et al., *DL4KG@ESWC*)
(Monnin, Raïssi, et al., *Semantic Web Journal*)

Thank you for your attention!



Bibliography i

- Tim Berners-Lee, James Hendler, and Ora Lassila (2001). "The semantic web". In: *Scientific american* 284.5, pp. 28–37.
- HongYun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang (2018). "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications". In: *IEEE Trans. Knowl. Data Eng.* 30.9, pp. 1616–1637. DOI: `10.1109/TKDE.2018.2807452`. URL: `https://doi.org/10.1109/TKDE.2018.2807452`.
- Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy (2020). "Machine Learning on Graphs: A Model and Comprehensive Taxonomy". In: *CoRR* abs/2005.03675. arXiv: `2005.03675`. URL: `https://arxiv.org/abs/2005.03675`.
- Nicholas Frosst, Nicolas Papernot, and Geoffrey E. Hinton (2019). "Analyzing and Improving Representations with the Soft Nearest Neighbor Loss". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2012–2020. URL: `http://proceedings.mlr.press/v97/frosst19a.html`.
- Ramanathan V. Guha (2015). "Towards A Model Theory for Distributed Representations". In: *2015 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 22-25, 2015*. AAAI Press. URL: `http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10220`.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu (2020). "A Survey on Knowledge Graphs: Representation, Acquisition and Applications". In: *CoRR* abs/2002.00388. arXiv: `2002.00388`. URL: `https://arxiv.org/abs/2002.00388`.
- Thomas N. Kipf and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: `https://openreview.net/forum?id=SJU4ayYgl`.
- Pierre Monnin, Clément Jonquet, Joël Legrand, Amedeo Napoli, and Adrien Coulet (2017). "PGxO: A very lite ontology to reconcile pharmacogenomic knowledge units". In: *Methods, tools & platforms for Personalized Medicine in the Big Data Era, NETTAB 2017*. Vol. 5. PeerJ PrePrints. PeerJ, e3140. DOI: `10.7287/peerj.preprints.3140v1`.

Bibliography ii

- Pierre Monnin, Joël Legrand, Graziella Husson, Patrice Ringot, Andon Tchechmedjiev, Clément Jonquet, Amedeo Napoli, and Adrien Coulet (2019). "PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison". In: *BMC Bioinformatics* 20-S.4, 139:1–139:16. DOI: `10.1186/s12859-019-2693-9`.
- Pierre Monnin, Amedeo Napoli, and Adrien Coulet (2017). "Discovering Subsumption Axioms with Concept Annotation". In: *Gestion de Données – Principes, Technologies et Applications (BDA 2017)*. URL: `https://hal.inria.fr/hal-01671454/document`.
- Pierre Monnin, Chedy Raïssi, Amedeo Napoli, and Adrien Coulet (2019). "Knowledge Reconciliation with Graph Convolutional Networks: Preliminary Results". In: *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located with the 16th Extended Semantic Web Conference 2019 (ESWC 2019), Portoroz, Slovenia, June 2, 2019*. Ed. by Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, Diego Reforgiato Recupero, and Harald Sack. Vol. 2377. CEUR Workshop Proceedings. CEUR-WS.org, pp. 47–56. URL: `http://ceur-ws.org/Vol-2377/paper_6.pdf`.
- Pierre Monnin, Chedy Raïssi, Amedeo Napoli, and Adrien Coulet (2022). "Discovering alignment relations with Graph Convolutional Networks: A biomedical case study". In: *Semantic Web* 13.3, pp. 379–398. DOI: `10.3233/SW-210452`.
- Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich (2016). "A Review of Relational Machine Learning for Knowledge Graphs". In: *Proceedings of the IEEE* 104.1, pp. 11–33. DOI: `10.1109/JPROC.2015.2483592`. URL: `https://doi.org/10.1109/JPROC.2015.2483592`.
- Natasha Noy, Alan Rector, Pat Hayes, and Chris Welty (2006). "Defining N-ary Relations on the Semantic Web". In: *W3C working group note* 12.4.

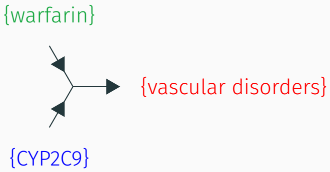
Bibliography iii

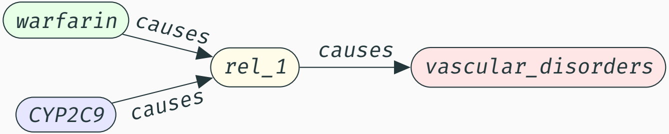
- Ning Pang, Weixin Zeng, Jiuyang Tang, Zhen Tan, and Xiang Zhao (2019). "Iterative Entity Alignment with Improved Neural Attribute Embedding". In: *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located with the 16th Extended Semantic Web Conference 2019 (ESWC 2019), Portoroz, Slovenia, June 2, 2019*. Ed. by Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, Diego Reforgiato Recupero, and Harald Sack. Vol. 2377. CEUR Workshop Proceedings. CEUR-WS.org, pp. 41–46. URL: http://ceur-ws.org/Vol-2377/paper_5.pdf.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling (2018). "Modeling Relational Data with Graph Convolutional Networks". In: *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*. Ed. by Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam. Vol. 10843. Lecture Notes in Computer Science. Springer, pp. 593–607. DOI: 10.1007/978-3-319-93417-4_38. URL: https://doi.org/10.1007/978-3-319-93417-4_38.
- Quan Wang, Zhendong Mao, Bin Wang, and Li Guo (2017). "Knowledge Graph Embedding: A Survey of Approaches and Applications". In: *IEEE Trans. Knowl. Data Eng.* 29.12, pp. 2724–2743. DOI: 10.1109/TKDE.2017.2754499. URL: https://doi.org/10.1109/TKDE.2017.2754499.
- Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang (2018). "Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Association for Computational Linguistics, pp. 349–357. DOI: 10.18653/v1/d18-1032. URL: https://doi.org/10.18653/v1/d18-1032.

Supplementary slides

PGxLOD: a knowledge graph for pharmacogenomics

- Scope: representing PGx *n*-ary relationships to reconcile and trace them
- Representation of *n*-ary relationships through **reification** (Noy et al. 2006)





Results of a first rule-based matching approach

		PharmGKB (sd)	PharmGKB (ca)	Literature	EHRs
Links from Rule 1 (=) Encoded by ***owl:sameAs***	PharmGKB (sd)	166	0	0	0
	PharmGKB (ca)	0	10,134	0	0
	Literature	0	0	122,646	0
	EHRs	0	0	0	0
Links from Rule 2 (∼) Encoded by ***skos:closeMatch***	PharmGKB (sd)	0	5	0	0
	PharmGKB (ca)	5	1,366	0	0
	Literature	0	0	16,692	0
	EHRs	0	0	0	0
Links from Rule 3 (≺) Encoded by ***skos:broadMatch***	PharmGKB (sd)	87	3	15	0
	PharmGKB (ca)	9,325	605	42	0
	Literature	0	0	75,138	0
	EHRs	0	0	0	0
Links from Rule 4 (≲) Encoded by ***skos:relatedMatch***	PharmGKB (sd)	20	0	0	0
	PharmGKB (ca)	0	110	0	0
	Literature	0	0	18,050	0
	EHRs	0	0	0	0
Links from Rule 5 ($\propto$) Encoded by ***skos:related*** similarity $\geq$ 0.8	PharmGKB (sd)	100,596	287,670	414	2
	PharmGKB (ca)	287,670	706,270	1,103	19
	Literature	414	1,103	1,082,074	15
	EHRs	2	19	15	0

sd: Structured Data
ca: Clinical Annotations

Results of a first rule-based matching approach

		PharmGKB (sd)	PharmGKB (ca)	Literature	EHRs
Links from Rule 1 (=) Encoded by ***owl:sameAs***	PharmGKB (sd)	166	0	0	0
	PharmGKB (ca)	0	10,134	0	0
	Literature	0	0	122,646	0
	EHRs	0	0	0	0
Links from Rule 2 ($\sim$) Encoded by ***skos:closeMatch***	PharmGKB (sd)	0	5	0	0
	PharmGKB (ca)	5	1,366	0	0
	Literature	0	0	16,692	0
	EHRs	0	0	0	0
Links from Rule 3 ($\preccurlyeq$) Encoded by ***skos:broadMatch***	PharmGKB (sd)	87	3	15	0
	PharmGKB (ca)	9,325	605	42	0
	Literature	0	0	75,138	0
	EHRs	0	0	0	0
Links from Rule 4 ($\lesssim$) Encoded by ***skos:relatedMatch***	PharmGKB (sd)	20	0	0	0
	PharmGKB (ca)	0	110	0	0
	Literature	0	0	18,050	0
	EHRs	0	0	0	0
Links from Rule 5 ($\propto$) Encoded by ***skos:related*** similarity $\geq$ 0.8	PharmGKB (sd)	100,596	287,670	414	2
	PharmGKB (ca)	287,670	706,270	1,103	19
	Literature	414	1,103	1,082,074	15
	EHRs	2	19	15	0

sd: Structured Data
ca: Clinical Annotations

Results of a first rule-based matching approach

		PharmGKB (sd)	PharmGKB (ca)	Literature	EHRs
Links from Rule 1 (=) Encoded by ***owl:sameAs***	PharmGKB (sd)	166	0	0	0
	PharmGKB (ca)	0	10,134	0	0
	Literature	0	0	122,646	0
	EHRs	0	0	0	0
Links from Rule 2 ($\sim$) Encoded by ***skos:closeMatch***	PharmGKB (sd)	0	5	0	0
	PharmGKB (ca)	5	1,366	0	0
	Literature	0	0	16,692	0
	EHRs	0	0	0	0
Links from Rule 3 ($\preccurlyeq$) Encoded by ***skos:broadMatch***	PharmGKB (sd)	87	3	15	0
	PharmGKB (ca)	9,325	605	42	0
	Literature	0	0	75,138	0
	EHRs	0	0	0	0
Links from Rule 4 ($\lessgtr$) Encoded by ***skos:relatedMatch***	PharmGKB (sd)	20	0	0	0
	PharmGKB (ca)	0	110	0	0
	Literature	0	0	18,050	0
	EHRs	0	0	0	0
Links from Rule 5 ($\propto$) Encoded by ***skos:related*** similarity $\geq$ 0.8	PharmGKB (sd)	100,596	287,670	414	2
	PharmGKB (ca)	287,670	706,270	1,103	19
	Literature	414	1,103	1,082,074	15
	EHRs	2	19	15	0

sd: Structured Data
ca: Clinical Annotations

Statistics of the considered neighborhood in PGxLOD

	# nodes	# edges	# predicates
PGxLOD	11,808,396	43,341,712	416
$\mathcal{G}_0$	3,758,814	39,956,844	689
$\mathcal{G}_1$	3,879,081	46,960,365	733
$\mathcal{G}_2$	3,758,814	22,085,701	347
$\mathcal{G}_3$	3,758,814	41,048,190	697
$\mathcal{G}_4$	3,758,928	42,691,984	701
$\mathcal{G}_5$	3,882,945	27,277,789	375

Learning embeddings, clustering, and evaluation

- All gold clusterings $\mathcal{C}_i$ with $\mathcal{G}_0$ (no inference rules) and $\mathcal{G}_5$ (all inference rules)
- All graphs $\mathcal{G}_j$ with gold clustering $\mathcal{C}_0$ (all alignment relations)
- 5-fold cross validation
- 3 layer GCN $\rightarrow$ 3-hop neighborhood considered
- Clustering on embeddings of nodes belonging to gold clusters

Algorithm	Parameter
Ward	# clusters to find
Single	# clusters to find
OPTICS	Min size of clusters

Metric	Abbr.	Domain
Unsupervised Clustering Accuracy	ACC	$[0, 1]$
Adjusted Rand Index	ARI	$[-1, 1]$
Normalized Mutual Information	NMI	$[0, 1]$

Clustering performance: all gold clusterings $\mathcal{C}_i$ with $\mathcal{G}_0$ and $\mathcal{G}_5$

		$\mathcal{G}_0$ (no inference rules)			$\mathcal{G}_5$ (all inference rules)		
		ACC	ARI	NMI	ACC	ARI	NMI
$\mathcal{C}_0$	Single	0.66	0.53	0.52	0.74	0.61	0.54
$\mathcal{C}_1$	Single	0.41	0.18	0.41	0.72	0.53	0.52
$\mathcal{C}_2$	Single	0.99	0.99	0.99	0.99	0.99	0.99
$\mathcal{C}_3$	Ward	0.92	0.90	0.94	0.86	0.81	0.89
$\mathcal{C}_4$	Ward	0.99	0.90	0.86	0.99	0.91	0.88
$\mathcal{C}_5$	Single	0.81	0.31	0.25	0.82	0.32	0.26
$\mathcal{C}_6$	Single	0.63	0.56	0.70	0.74	0.76	0.76

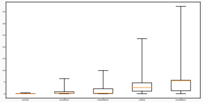
- $\mathcal{C}_0$ and $\mathcal{C}_1$
  - Mix different alignment relations $\rightarrow$ Difficult task
  - Consistent improvement w/ $\mathcal{G}_5$
- $\mathcal{C}_2$, $\mathcal{C}_3$, $\mathcal{C}_4$, $\mathcal{C}_5$, and $\mathcal{C}_6$
  - Only one alignment relation each
  - No homogeneous improvement

Clustering performance: all graphs $\mathcal{G}_j$ with gold clustering $\mathcal{C}_0$

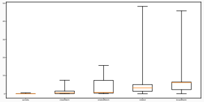
Graph	ACC	ARI	NMI	Performance
$\mathcal{G}_0$ (no inference rules)	0.66	0.53	0.52	Baseline
$\mathcal{G}_1$ (`sameAs` contraction)	0.73	0.78	0.51	Improvements
$\mathcal{G}_2$ (inverses and symmetry of predicates)	0.62	0.47	0.48	Deterioration
$\mathcal{G}_3$ (hierarchy of predicates)	0.70	0.58	0.52	Improvements
$\mathcal{G}_4$ (hierarchy of classes)	0.56	0.42	0.50	Consistent deterioration
$\mathcal{G}_5$ (all inference rules)	0.74	0.61	0.54	Improvements – Best results

Results on algorithm w/ best performance: Single

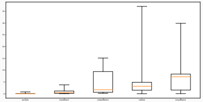
Distance analysis: rediscovery of alignment relations?



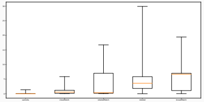
(a) $\mathcal{G}_0$ – Fold 1



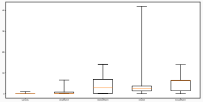
(b) $\mathcal{G}_0$ – Fold 2



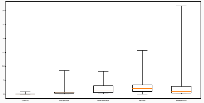
(c) $\mathcal{G}_0$ – Fold 3



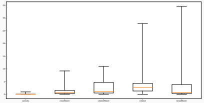
(d) $\mathcal{G}_0$ – Fold 4



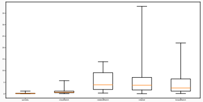
(e) $\mathcal{G}_0$ – Fold 5



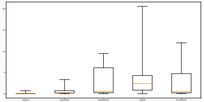
(f) $\mathcal{G}_5$ – Fold 1



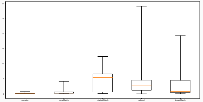
(g) $\mathcal{G}_5$ – Fold 2



(h) $\mathcal{G}_5$ – Fold 3



(i) $\mathcal{G}_5$ – Fold 4

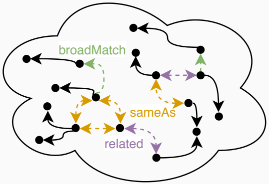


(j) $\mathcal{G}_5$ – Fold 5

In each figure, links are from left to right:

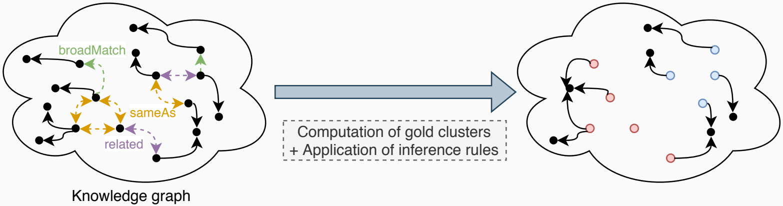
`owl:sameAs`, `skos:closeMatch`, `skos:relatedMatch`, `skos:related`, `skos:broadMatch`

Approach outline



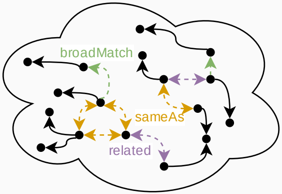
Knowledge graph

Approach outline

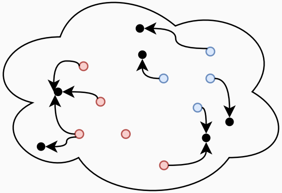


Gold clusterings: from similarity links to gold clusters

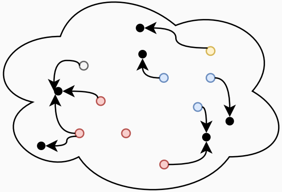
	Rule 1 (=) *owl:sameAs*	Rule 2 ($\sim$) *skos:closeMatch*	Rule 4 ($\lessgtr$) *skos:relatedMatch*	Rule 5 ($\propto$) *skos:related*	Rule 3 ($\preccurlyeq$) *skos:broadMatch*
$\mathcal{C}_0$	×	×	×	×	×
$\mathcal{C}_1$	×	×	×	×	
$\mathcal{C}_2$	×				
$\mathcal{C}_3$		×			
$\mathcal{C}_4$			×		
$\mathcal{C}_5$				×	
$\mathcal{C}_6$					×



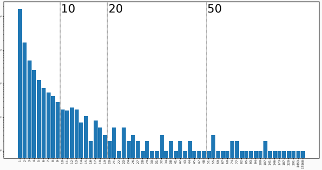
Initial KG



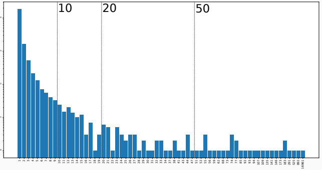
$\mathcal{C}_0$



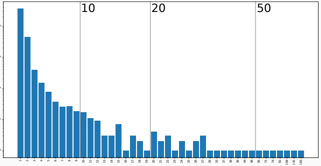
$\mathcal{C}_1$



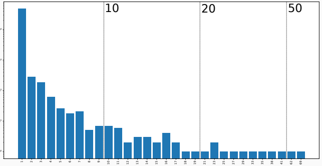
(a) $\mathcal{C}_0$ (max $= 17,568$)



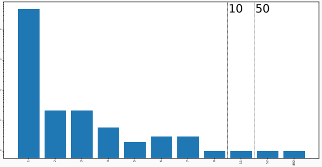
(b) $\mathcal{C}_1$ (max $= 16,961$)



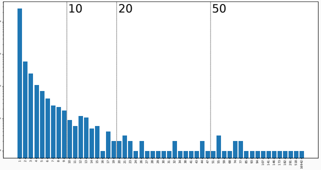
(c) $\mathcal{C}_2$ (max $= 183$)



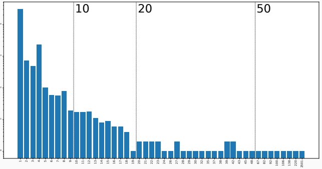
(d) $\mathcal{C}_3$ (max $= 69$)



(e) $\mathcal{C}_4$ (max $= 892$)

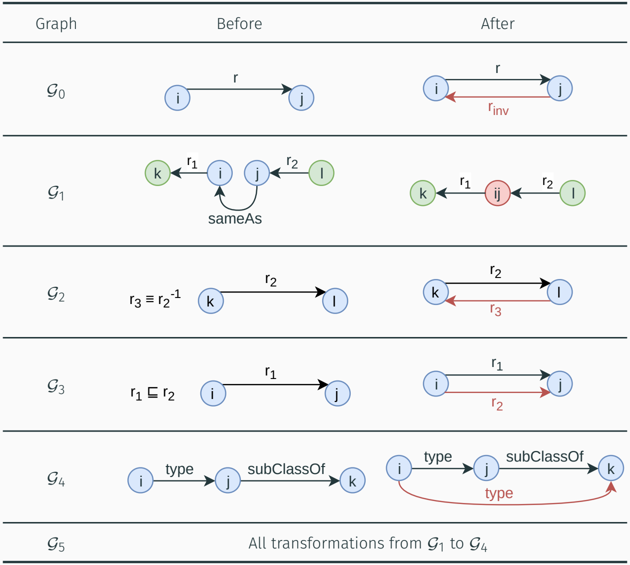


(f) $\mathcal{C}_5$ (max $= 16,942$)

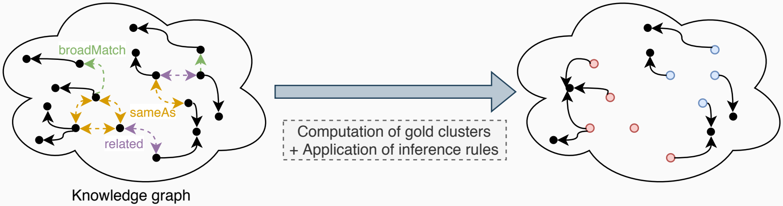


(g) $\mathcal{C}_6$ (max $= 2,501$)

Applying inference rules associated with domain knowledge



Approach outline



Approach outline