Supplementary material for SEKA: Seeking Knowledge Graph Anomalies

Asara Senaratne, Pouya Ghiasnezhad Omran, Peter Christen, Graham Williams

School of Computing, The Australian National University, Canberra ACT 2600 Australia.
asara.senaratn@anu.edu.au, p.g.omran@anu.edu.au, peter.christen@anu.edu.au, graham.williams@anu.edu.au

Our Approach Visualized

Following recent work in anomaly detection in Knowledge Graphs (KGs) (Senaratne et al. 2021; Zheng et al. 2019; Jia et al. 2018), our aim is to discover abnormal triples and entities that provide contradicting, semantically incorrect, and redundant information, or that contain invalid, incomplete, or missing information. Figure 1 provides a visualization of SEKA.

Knowledge Graph Definition

We consider a directed edge-labelled KG, $G = (V, E)$ containing a set of nodes (or vertices) V and a set of labelled edges E connecting these vertices. The Resource Description Framework (RDF) is a standardised data model based on directed edge-labelled graphs[1]. The RDF model defines three types of nodes in a graph: (1) Internationalized Resource Identifiers (IRIs) I which assigns a global identifier for entities I_e and relations I_r on the web (where $I = I_e \cup I_r$); (2) literals L which represents strings and other datatype values; and (3) blank nodes B which are anonymous nodes (not a URI reference or a literal) that do not have an identifier (Hogan et al. 2021). We therefore have the node set $V = (I_e \cup L \cup B)$ and edge set $E \in V \times I_r \times V$. Each edge $t \in E \in G$ is considered as a RDF triple. A triple (also named as a triplet) contains the three elements subject $s \in S$, predicate $p \in P$, and object $o \in O$. A triple is denoted as (s, p, o) where $(s, o) \in V$, and $(s \times p \times o) \in E$. Furthermore, $s \in (I, B)$, $p \in I_r$, and $o \in (I, L, B)$.

Synthetic Anomalies

As the four KGs we use for experimental evaluation do not contain labelled data, we manually inject anomalies. To achieve this, we introduce INK (INject anomalies to Knowledge graphs), which injects the following types of anomalies to a KG.

- Change each of subject, object, and predicate one at a time in a triple while preserving the type. For example, replace predicates used for a person with a predicate that is also used for a person. For example, (personA, isMarriedTo, personB) changed to (personA, hasChild, personB).
- Change the entity type of each of subject, object, and predicate one at a time in a triple. For example, replace predicates used for a person with a predicate that is not used for a person. For example, (personA, isMarriedTo, personB) changed to (personA, livesIn, personB).

[1]https://www.w3.org/TR/rdf-concepts/\#section-Graph-URIref

Dataset Statistics

Table 1 provides a summary of the sizes of the four real-world KGs that we used to conduct the experiments described in the main abstract. These KGs demonstrate the applicability of our approach on KGs from different domains of varying sizes.

Table 1: KG summaries.

KG	Entity count	Triple count where the object is a literal	Triple count where the object is an entity
YAGO-1	2,215,094	21,337,521	922,741
KBpedia	62,796	534,032	227,060
Wikidata	14,036,475	53,541,372	51,559,889
DSKG	5,952	22,202	828,086

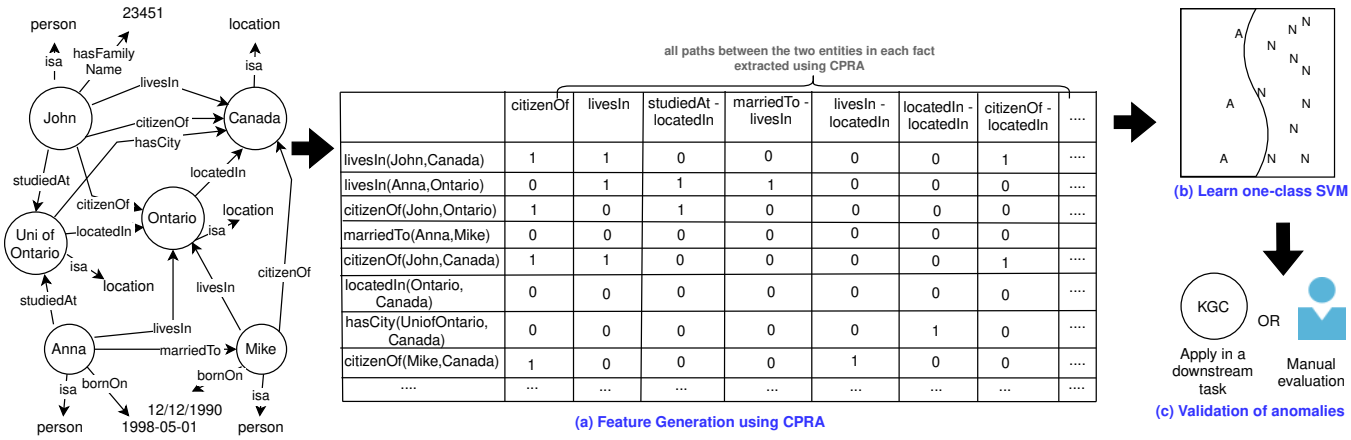
Additional Results

We used the two measures precision and recall to evaluate SEKA in performing fact anomaly detection on the four experimental datasets. We conducted the evaluation under three different percentages of anomalies injected in to the KGs.

As the results in Table 2 show, our approach performs well in identifying anomalies in all four KGs under all three anomaly percentages. SEKA performs the best in DSKG with 10% and 20% anomalies, while it performs the best in YAGO-1 with 30% anomalies.

References

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G. D.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; et al. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4): 1–37.



	citizenOf	livesIn	studiedAt - locatedIn	marriedTo - livesIn	livesIn - locatedIn	locatedIn - locatedIn	citizenOf - locatedIn	....
livesIn(John,Canada)	1	1	0	0	0	0	1	....
livesIn(Anna,Ontario)	0	1	1	1	0	0	0	....
citizenOf(John,Ontario)	1	0	1	0	0	0	0	....
marriedTo(Anna,Mike)	0	0	0	0	0	0	0	
citizenOf(John,Canada)	1	1	0	0	0	0	1	....
locatedIn(Ontario, Canada)	0	0	0	0	0	0	0	....
hasCity(UniofOntario, Canada)	0	0	0	0	0	1	0	....
citizenOf(Mike,Canada)	1	0	0	0	1	0	0	....
....	...	...	...	...	...	...	...	....

Figure 1: Overview of SEKA, our anomaly detection process to identify anomalous triples in a KG.

Table 2: Results obtained by SEKA for general fact anomaly detection with 10%, 20% and 30% anomalies in each KG.

KG	10% anomalies			20% anomalies			30% anomalies		
	Anomaly count	Precision	Recall	Anomaly count	Precision	Recall	Anomaly count	Precision	Recall
YAGO-1	92,274	0.99	0.99	184,548	0.90	0.89	276,822	0.86	0.87
KBpedia	22,706	0.82	0.81	45,412	0.70	0.70	68,118	0.72	0.72
Wikidata	5,155,989	0.81	0.80	10,311,978	0.79	0.78	15,467,967	0.72	0.71
DSKG	82,809	0.99	0.99	165,617	0.95	0.94	248,425	0.82	0.81

Jia, B.; Dong, C.; Chen, Z.; Chang, K.-C.; Sullivan, N.; and Chen, G. 2018. Pattern Discovery and Anomaly Detection via Knowledge Graph. In *International Conference on Information Fusion*, 2392–2399. IEEE.

Senaratne, A.; Omran, P. G.; Williams, G.; and Christen, P. 2021. Unsupervised Anomaly Detection in Knowledge Graphs. In *International Joint Conference on Knowledge Graphs*, 161–165. New York, USA: ACM.

Zheng, L.; Li, Z.; Li, J.; Li, Z.; and Gao, J. 2019. AddGraph: Anomaly Detection in Dynamic Graph Using Attention-based Temporal GCN. In *IJCAI*, 4419–4425. USA: IJCAI.